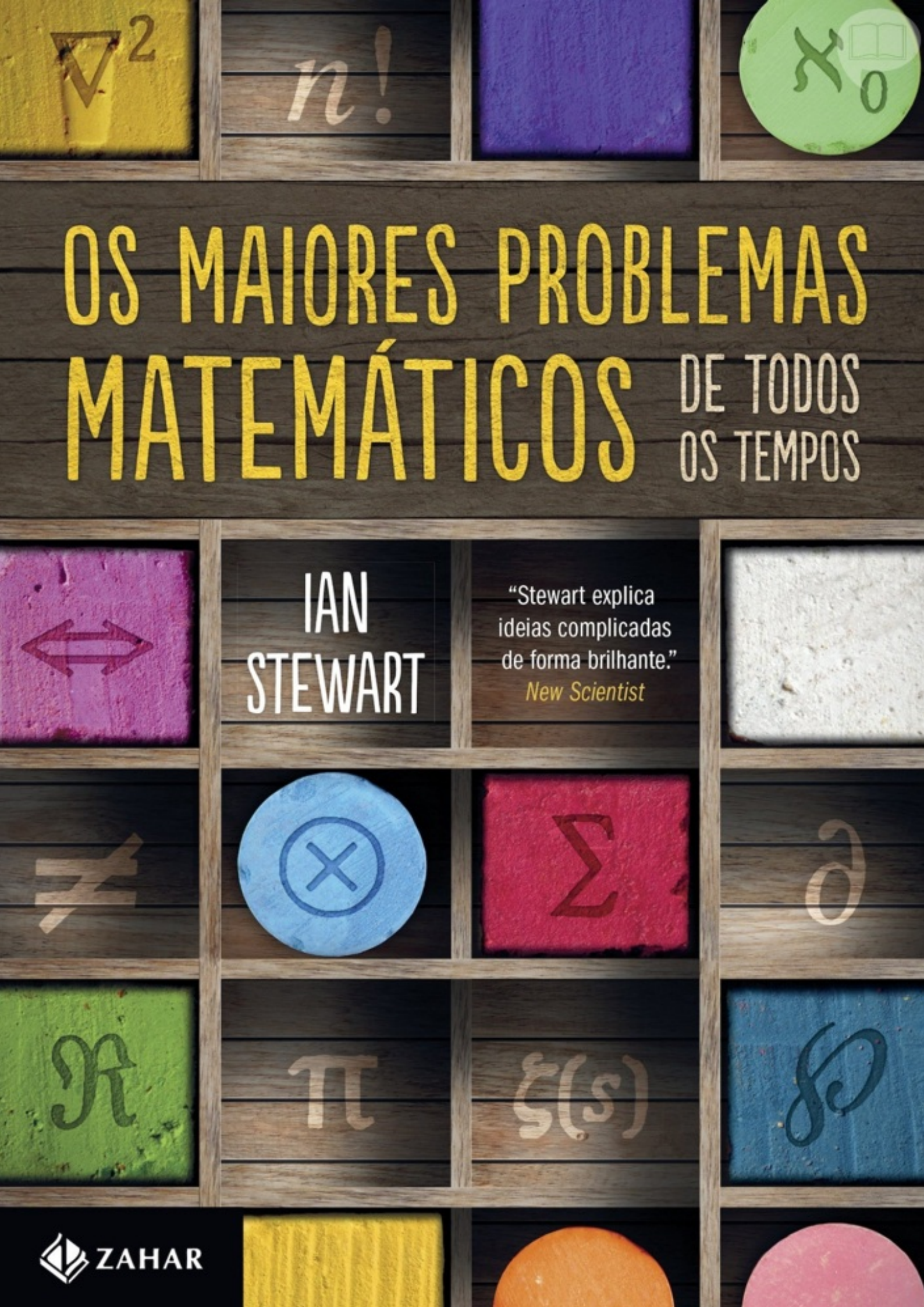
OS MAIORES PROBLEMAS MATEMÁTICOS DE TODOS OS TEMPOS

IAN STEWART

"Stewart explica ideias complicadas de forma brilhante."
New Scientist



ZAHAR

Uma história da matemática acessível, envolvente e esclarecedora contada através de 14 de seus maiores problemas

Com sua clareza habitual, o aclamado professor Ian Stewart, autor de sucessos como Almanaque das curiosidades matemáticas, apresenta um panorama fascinante das questões realmente grandiosas e desafiadoras da matemática, ao longo de três milênios de história.

Do último teorema de Fermat, que levou mais de 350 anos para ser solucionado, à complexa hipótese de Riemann, o Santo Graal da matemática, impenetrável após 150 anos.

O autor os explica em detalhes, descrevendo sua origem, importância e o contexto em que surgiram. De forma simples e objetiva, mostra ainda como esses grandes enigmas guiaram e continuam a instigar as mentes mais brilhantes do mundo, conduzindo a empreitada matemática por rumos radicalmente novos.

Cada capítulo é dedicado a um problema. Entre eles:
- A quadratura do círculo
- O teorema das quatro cores
- A conjectura de Kepler
- O último teorema de Fermat
- O problema dos três corpos
- A hipótese do mass gap

Nós precisamos saber. Nós havemos de saber.

DAVID HILBERT

(*Discurso sobre problemas matemáticos em 1930, por ocasião de sua cidadania honorária de Königsberg.*)[1]

Prefácio

A MATEMÁTICA É UM TEMA AMPLO, em perene crescimento, em perene mudança. Entre as inúmeras perguntas que os matemáticos se fazem e, em sua maior parte, respondem, algumas se destacam do restante: são picos proeminentes que pairam sobre os humildes sopés dos morros. Essas são as questões realmente grandes, os problemas difíceis e desafiadores que qualquer matemático daria o braço direito para resolver. Algumas permaneceram sem resposta por décadas, outras por séculos, algumas por milênios. Há ainda as que estão por ser conquistadas. O último teorema de Fermat permaneceu um enigma durante 350 anos, até que Andrew Wiles o desvendou após sete anos de labuta. A conjectura de Poincaré ficou em aberto por mais de um século até que foi solucionada por Grigori Perelman, gênio excêntrico que declinou de todas as honrarias acadêmicas e de um prêmio de 1 milhão de dólares pelo seu trabalho. A hipótese de Riemann continua a assombrar os matemáticos do mundo, impenetrável como sempre após 150 anos.

Os maiores problemas matemáticos de todos os tempos contém uma seleção das questões realmente grandiosas que têm conduzido a empreitada matemática por rumos radicalmente novos. O livro descreve suas origens, explica por que são importantes e as coloca no contexto da matemática e da ciência como um todo. Estão incluídos problemas resolvidos e não resolvidos, que abrangem mais de 2 mil anos de desenvolvimento matemático, mas seu foco principal está em questões que, ou permanecem abertas até hoje, ou foram solucionadas ao longo dos últimos cinquenta anos.

Um dos objetivos básicos da matemática é revelar a simplicidade subjacente de questões aparentemente complicadas. No entanto, ela pode nem sempre ser visível, porque a concepção matemática de “simples” apoia-se em muitos conceitos técnicos e difíceis. Uma característica importante deste livro é enfatizar simplicidades profundas e evitar – ou pelo menos explicar em termos diretos – as complexidades.

A MATEMÁTICA É MAIS RECENTE, e mais diversificada, do que a maioria de nós imagina. Numa estimativa aproximada, há cerca de 100 mil pesquisadores matemáticos no mundo, que produzem mais de *2 milhões* de páginas de matemática nova todo ano. Não "novos números", o que não é realmente a meta do empreendimento. Nem "novas somas" como as já existentes, porém maiores – embora façamos, sim, cálculos com somas bastante grandes. Um recente problema de álgebra, abordado por uma equipe com cerca de 25 matemáticos, foi descrito como "um cálculo do tamanho de Manhattan". Isso não era bem verdade, mas o erro foi ser uma comparação conservadora demais. A *resposta* era do tamanho de Manhattan; o cálculo era bem maior. Isso é impressionante, mas o que importa é a qualidade, não a quantidade. O cálculo do tamanho de Manhattan insere-se em ambas as classificações, porque fornece uma informação básica valiosa a respeito de um grupo de simetria que parece ser importante em física quântica, e é decididamente importante em matemática. A matemática brilhante pode ocupar uma linha, ou uma enciclopédia – o que o problema exigir.

Quando pensamos em matemática, o que nos vem logo à mente são páginas intermináveis de densos símbolos e fórmulas. Contudo, esses 2 milhões de páginas geralmente contêm mais palavras do que símbolos. As palavras estão ali para explicar os antecedentes do problema, o fluxo do argumento, o significado dos cálculos e como tudo isso se encaixa no sempre crescente edifício da matemática. Como comentou o grande Carl Friedrich Gauss por volta de 1800, a essência da matemática são "noções, não notações". Ideias, não símbolos. Mesmo assim, a linguagem usual para expressar ideias matemáticas é simbólica. Muitos artigos de pesquisa publicados contêm mais símbolos do que palavras. As fórmulas possuem uma precisão que nem sempre as palavras conseguem acompanhar.

No entanto, muitas vezes é possível explicar as ideias deixando de lado a maioria dos símbolos. *Os maiores problemas matemáticos de todos os tempos* toma isso como diretriz. O livro ilumina o que os matemáticos fazem, como pensam, e por que seu assunto é interessante e importante. Significativamente, ele mostra como os matemáticos de hoje estão se erguendo para desafios colocados pelos seus predecessores, à medida que um a um os grandes enigmas do passado vão se rendendo às poderosas técnicas do presente, transformando a matemática e a ciência do futuro. A matemática encaixa-se entre as grandes conquistas da humanidade, e seus maiores problemas – solucionados ou não – têm guiado e estimulado seu espantoso poder por milênios, tanto no passado quanto no que está por vir.

Coventry, junho de 2012

1. Grandes problemas

OS PROGRAMAS DE TELEVISÃO sobre matemática são raros, e os bons mais raros ainda. Um dos melhores, em termos de interesse e envolvimento da audiência, bem como de conteúdo, foi *O último teorema de Fermat*. O programa foi produzido em 1996 por John Lynch para a série *Horizo*n, carro-chefe de ciência popular da BBC. Simon Singh, que também esteve envolvido na realização, transformou a história em um livro espetacular que se tornou best-seller.[1] Em um site da internet, ele ressaltou que o extraordinário sucesso do programa foi uma surpresa:

Foram cinquenta minutos de matemáticos falando sobre matemática, o que não é a receita óbvia para um sucesso de audiência na TV, mas o resultado foi um programa que capturou a imaginação do público e recebeu aclamação da crítica. Ganhou o prêmio Bafta[a] como melhor documentário, um Priz Italia, outros prêmios internacionais e uma indicação para o Emmy; isso prova que a matemática pode ser tão emocionante e tão cativante como qualquer outro assunto do planeta.

Penso que existem diversos motivos para o sucesso tanto do programa de televisão como do livro, e esses motivos têm implicações para a história que quero contar aqui. Para manter o foco da discussão, vou me concentrar no documentário de TV.

O último teorema de Fermat é, na verdade, um dos grandes problemas matemáticos, surgido de um comentário aparentemente inócuo que um dos mais importantes matemáticos do século XVII escreveu na margem de um livro-texto clássico. O problema tornou-se notório porque ninguém conseguiu provar o que afirmava a nota escrita por Pierre de Fermat, permanecendo assim por mais de trezentos anos, apesar dos árduos esforços de pessoas com inteligência extraordinária. Então, quando o matemático britânico Andrew Wiles finalmente decifrou o problema em 1995, a magnitude de sua façanha ficou óbvia para qualquer um. Nem sequer precisávamos saber qual era o problema, muito menos como fora resolvido. Tratava-se do equivalente matemático da primeira escalada do

monte Everest.

Além do seu significado para a matemática, a solução de Wiles envolvia também uma densa história de interesse humano. Aos dez anos de idade, ele ficara tão intrigado pelo problema que decidiu tornar-se matemático para resolvê-lo. Executou a primeira parte do plano, e chegou a se especializar em teoria dos números, a área geral à qual pertence o teorema de Fermat. Porém, quanto mais aprendia matemática real, mais impossível toda a empreitada parecia. O último teorema de Fermat era uma espantosa curiosidade, uma questão isolada do tipo que qualquer teórico dos números poderia elaborar em sonho sem um fiapo de evidência convincente. Ele não se encaixava em nenhuma estrutura consistente de técnica. Em uma carta para Heinrich Olbers, o grande Gauss o desdenhara como inoportuno, dizendo que o problema tinha "pouco interesse para mim, já que uma enorme quantidade de tais proposições, que não podem ser provadas nem refutadas, pode ser formulada com facilidade".[2] Wiles decidiu que seu sonho de infância fora irrealista, e guardou Fermat no fundo do armário. Mas então, como por um milagre, outros matemáticos subitamente fizeram uma descoberta que ligava o problema a um tópico central na teoria dos números, algo no qual Wiles já era perito. Gauss, contrariando sua característica, subestimara a significação do problema, e não tinha consciência de que ele podia ser ligado a uma área profunda, ainda que aparentemente não relacionada, da matemática.

Com esse elo estabelecido, Wiles pôde agora trabalhar no enigma de Fermat e ao mesmo tempo fazer pesquisa digna de crédito na moderna teoria dos números. Melhor ainda, se Fermat não desse certo, qualquer coisa significativa que ele descobrisse ao tentar provar o teorema seria digna de ser publicada. Assim, Fermat foi resgatado do fundo do armário, e Wiles começou a pensar a sério no problema. Após sete anos de obsessiva pesquisa, levada a cabo de modo particular e em segredo – uma precaução inusitada em matemática –, convenceu-se de que havia achado uma solução. Proferiu uma série de palestras numa prestigiosa conferência sobre teoria dos números, com um título obscuro que não enganou ninguém.[3] A empolgante novidade estourou, tanto na mídia como no mundo acadêmico: o último teorema de Fermat havia sido provado.

A prova era impressionante e elegante, repleta de boas ideias. Infelizmente, os peritos logo descobriram uma séria lacuna em sua lógica. Nas tentativas de demolir grandes problemas não resolvidos da matemática, esse tipo de acontecimento é deprimentemente comum, e

quase sempre se mostra fatal. Todavia, dessa vez as Moiras foram gentis. Com a ajuda de seu ex-aluno Richard Taylor, Wiles conseguiu preencher a lacuna, reparar a prova e completar sua solução. A carga emocional envolvida ficou vividamente clara no programa de televisão: deve ter sido a única ocasião em que um matemático irrompeu em lágrimas na tela, só de lembrar os eventos traumáticos e o triunfo final.

Você deve ter notado que eu não revelei o que *é* o último teorema de Fermat. Isso é proposital; ele será abordado em seu devido lugar. Até onde vai o sucesso do programa de TV, realmente não conta. Na verdade, os matemáticos nunca se importaram muito se o teorema que Fermat rabiscou na margem é verdadeiro ou falso, porque nada de muito importante está ligado à resposta. Então, por que todo o alvoroço? Porque muitíssima coisa está relacionada à incapacidade da comunidade matemática de *encontrar* a resposta. Não é apenas um golpe na nossa autoestima: significa que as teorias matemáticas existentes estão perdendo algo vital. Além disso, o teorema é muito fácil de formular, o que contribui para o seu ar de mistério. Como pode algo que parece tão simples revelar-se tão difícil?

Embora os matemáticos realmente não se importassem com a resposta, importavam-se profundamente em não saber qual era ela. E interessavam-se ainda mais em achar um método que pudesse solucionar o problema, porque ele certamente lançaria luz não só sobre a questão de Fermat, mas sobre várias outras indagações. Isso é o que geralmente ocorre com grandes problemas matemáticos: são os métodos usados para resolvê-los, mais do que os resultados em si, o que mais importa. É claro que às vezes o resultado em si também importa: depende de quais são suas consequências.

A solução de Wiles é complicada e técnica demais para a televisão; na verdade, os detalhes são acessíveis apenas para especialistas.[4] Mas a prova envolve uma bela história matemática, como veremos na devida hora, porém qualquer tentativa de explicá-la na TV teria feito perder imediatamente a maior parte da audiência. Em vez disso, o programa concentrou-se sensatamente em uma pergunta mais pessoal: como é enfrentar um problema matemático, notoriamente difícil, que carrega uma pesada bagagem histórica? Mostrou-se aos telespectadores que existia um pequeno mas dedicado grupo de matemáticos, espalhados pelo mundo, que se importava profundamente com sua área de pesquisa, conversando entre si, tomando notas do trabalho uns dos outros e dedicando grande parte de

suas vidas ao progresso do conhecimento matemático. Seu investimento emocional e interação social foram vividamente mostrados. Não eram autômatos inteligentes, mas gente de verdade, engajada no seu tema. Essa foi a mensagem.

Esses são os três grandes motivos de o programa ter sido um sucesso: um problema de primeira grandeza, um herói com uma admirável história humana e um elenco de apoio composto de pessoas envolvidas emocionalmente. Mas suspeito que havia um quarto motivo, não tão notável. A maioria dos não matemáticos raramente ouve falar a respeito de novos desenvolvimentos nessa área, por uma variedade de razões perfeitamente sensatas: de qualquer maneira, não estão nem um pouco interessados; os jornais dificilmente mencionam algo sobre matemática; quando o fazem, muitas vezes é em tom jocoso ou trivial; e no cotidiano quase nada parece ser afetado por qualquer coisa que os matemáticos estejam fazendo nos bastidores. Com muita frequência, a matemática escolar é apresentada num livro fechado, no qual para toda pergunta há uma resposta. Os estudantes podem facilmente vir a imaginar que matemática nova é tão raro como dentes em uma galinha.

Desse ponto de vista, a grande notícia não era a de que o último teorema de Fermat havia sido provado, e sim a de que finalmente *alguém tinha feito matemática nova*. Considerando que os matemáticos tinham levado mais de trezentos anos para achar uma solução, muitos espectadores inconscientemente concluíram que essa arrancada era a primeira matemática nova importante descoberta nos últimos três séculos. Não estou sugerindo que tenham acreditado nisso *explicitamente*. Essa deixa de ser uma posição sustentável no instante em que seja feita alguma pergunta óbvia do tipo: “Por que o governo gasta tanto dinheiro nos departamentos de matemática das universidades?” Mas, de maneira inconsciente, era uma premissa comum de omissão, não questionada e não examinada. E fazia a magnitude da façanha de Wiles parecer ainda maior.

Um dos objetivos deste livro é mostrar que a pesquisa matemática está prosperando, com novas descobertas sendo feitas o tempo todo. Não se ouve falar muito dessa atividade porque a maior parte é técnica demais para não especialistas, porque a maior parte da mídia é cautelosa e desconfiada em relação a algo mais desafiador do que *The X Factor*, e porque as aplicações da matemática são deliberadamente ocultas para evitar causar alarme. “O quê? Meu iPhone depende de matemática avançada? Como é que vou entrar no Facebook se fui reprovado nos

exames de matemática?”

HISTORICAMENTE, uma matemática nova costuma surgir a partir de descobertas em outras áreas. Quando Isaac Newton deduziu suas leis do movimento e sua lei da gravidade, que juntas descrevem o movimento dos planetas, ele não lapidou o problema da compreensão do sistema solar. Ao contrário, os matemáticos tiveram de se atracar com toda uma gama nova de questões: sim, sabemos as leis, mas em que elas implicam? Newton inventou o cálculo para responder a essa pergunta, mas seu novo método também tem limitações. Muitas vezes ele reformula a pergunta em outros termos em vez de fornecer uma resposta, transformando o problema num tipo especial de fórmula, chamada equação diferencial, cuja *solução* é a resposta. Mas ainda é preciso resolver a equação. Não obstante, o cálculo foi um brilhante começo. Mostrou-nos que as respostas eram possíveis, e forneceu um meio efetivo de buscá-las, que continua a proporcionar importantes percepções mais de trezentos anos depois.

À medida que o conhecimento matemático coletivo da humanidade foi aumentando, uma segunda fonte de inspiração passou a desempenhar um papel crescente na criação de ainda mais: as exigências internas da própria matemática. Se, por exemplo, você sabe como resolver equações algébricas de primeiro, segundo, terceiro e quarto graus, então não é preciso muita imaginação para se perguntar acerca do quinto grau. (O grau é basicamente uma medida de complexidade, mas você nem precisa saber o que é para fazer-se a pergunta óbvia.) Se uma solução mostra-se arredia, como aconteceu, esse fato *em si* faz com que os matemáticos fiquem ainda mais determinados a encontrar uma resposta, tenha ou não o resultado aplicações úteis.

Não estou sugerindo que aplicações não tenham importância. Mas se um elemento matemático específico continua aparecendo em questões sobre a física das ondas – ondas do mar, vibrações, som, luz –, então seguramente faz sentido investigar o dispositivo em si. Você não necessita saber de antemão exatamente como qualquer ideia original será usada: o tópico das ondas é comum a tantas áreas importantes que novas percepções significativas estão propensas a ser úteis para alguma coisa. Nesse caso, esta alguma coisa inclui rádio, televisão e radar.[5] Se alguém pensa numa maneira diferente de entender o fluxo de calor e surge com uma brilhante técnica original, que infelizmente careça de sustentação

matemática, então faz sentido resolver a coisa toda como *um elemento de matemática*. Mesmo que você não dê importância a como o calor flui, os resultados podem muito bem ser aplicados em outra parte. A análise de Fourier, que surgiu dessa linha de investigação particular, é sem dúvida a ideia matemática isolada mais útil já encontrada. Ela escora as modernas telecomunicações, possibilita a existência de câmeras digitais, ajuda a limpar gravações e filmes antigos, e uma extensão moderna é usada pelo FBI para arquivar registros de impressões digitais.[6]

Após alguns milhares de anos com esse tipo de intercâmbio entre as aplicações externas da matemática e sua estrutura interna, esses dois aspectos do assunto tornaram-se tão densamente entrelaçados que é quase impossível separá-los. Entretanto, as atitudes mentais envolvidas são mais fáceis de serem distinguidas, levando a uma classificação ampla da matemática em dois tipos: pura e aplicada. Essa divisão é justificável como um meio rápido e aproximado de situar as ideias matemáticas no panorama intelectual, mas não é uma descrição extremamente precisa da área em si. Na melhor das hipóteses, ela distingue duas extremidades de um espectro contínuo de estilos matemáticos. Na pior, ela desvirtua quais as partes da área são úteis e de onde vêm as ideias. Como em todos os ramos da ciência, o que confere à matemática o seu poder é a *combinação* de raciocínio abstrato e a inspiração do mundo exterior, cada uma alimentando a outra. Não só é impossível separar as duas partes: é totalmente sem sentido.

A maioria dos problemas matemáticos realmente importantes, os grandes problemas dos quais este livro trata, surgiram no campo matemático mediante uma espécie de atitude intelectual de "observar o próprio umbigo". A razão é simples: são problemas *matemáticos*. A matemática muitas vezes parece uma coleção de áreas isoladas, cada uma com suas técnicas especiais peculiares: álgebra, análise, geometria, trigonometria, combinatória, probabilidade. Ela tende a ser ensinada dessa forma, por um bom motivo: situar cada tópico separadamente, em uma área específica bem-definida, ajuda o aluno a organizar o material em sua cabeça. É uma primeira aproximação razoável para a estrutura da matemática, sobretudo aquela que foi estabelecida há muito tempo. Nas fronteiras da pesquisa, porém, essa delineação clara geralmente tende a se romper. Não apenas pela razão de que as fronteiras entre as principais áreas da matemática ficam obscuras, e sim porque elas não existem de fato.

Todo matemático dedicado à pesquisa tem consciência de que, a

qualquer momento, de forma súbita e imprevisível, o problema no qual está trabalhando pode requerer ideias de uma área aparentemente não correlacionada. Na verdade, um novo estudo muitas vezes combina áreas. Por exemplo, a maior parte da minha própria pesquisa é centrada na formação de padrões em sistemas dinâmicos, que mudam com o tempo segundo regras específicas. Um exemplo típico é a maneira como os animais se movem. Um cavalo trotando repete a mesma sequência de movimentos das patas, e existe um padrão claro: as patas tocam o chão juntas em pares relacionados diagonalmente. Isto é, primeiro a dianteira esquerda junto com a traseira direita; depois as outras duas. Será esse um problema sobre padrões, e nesse caso os métodos apropriados provêm da teoria dos grupos, a álgebra da simetria? Ou será um problema sobre dinâmica, sendo então a área apropriada a das equações diferenciais ao estilo newtoniano?

A resposta, por definição, é que ele precisa ser as duas coisas, e não a sua interseção, que é aquilo que ambas as áreas têm em comum – ou seja, basicamente nada. Em vez disso, é uma "área" nova, que abarca duas das divisões tradicionais da matemática. É como uma ponte sobre um rio que separa dois países: ela liga os dois, mas não pertence a nenhum deles. Mas essa ponte não é uma estreita faixa de estrada, ela tem tamanho comparável a cada um dos países. Mais importante ainda, os métodos envolvidos não se limitam a essas duas áreas. Na verdade, praticamente todo curso de matemática que um dia frequentei desempenhou um papel em algum ponto da minha pesquisa. O curso sobre a teoria de Galois na minha graduação em Cambridge tratava de como resolver (mais precisamente, por que não podemos resolver) uma equação algébrica de quinto grau. Aquele a respeito da teoria dos grafos tratava de pontos (vértices) ligados por linhas (arestas) formando configurações (grafos). Nunca fiz um curso de sistemas dinâmicos, pois meu doutorado foi em álgebra, mas com o passar dos anos fui pegando a base, de estados estáveis até o caos. Teoria de Galois, teoria dos gráficos, sistemas dinâmicos: três áreas distintas. Ou assim julgava eu até 2011, quando quis compreender como detectar uma dinâmica caótica em uma rede de sistemas dinâmicos, e um passo crucial dependia das coisas que aprendi 45 anos antes no curso de teoria de Galois.

A matemática, portanto, não é como um mapa político do mundo, com cada especialidade ordenadamente cercada por uma fronteira clara, cada país nitidamente distinguido de seus vizinhos por verde, rosa ou azul-claro.

É mais como uma paisagem natural, onde nunca se pode realmente saber onde termina um vale e começa o morro, onde a floresta se funde com o bosque, os arbustos e a planície gramada, onde os lagos inserem regiões de água em todo outro tipo de terreno, onde os rios ligam as encostas cobertas de gelo das montanhas com os oceanos distantes. Mas essa paisagem matemática em constante mudança não consiste de rochas, águas e plantas, mas de ideias; ela é unida não pela geografia, mas pela lógica. E é uma paisagem dinâmica, que muda à medida que novas ideias e novos métodos vão sendo descobertos ou inventados. Conceitos importantes com amplas implicações são como picos montanhosos, técnicas com inúmeras utilidades são como rios largos que transportam viajantes por planícies férteis. Quanto mais claramente definida torna-se a paisagem, mais fácil é localizar picos não escalados, ou terrenos não explorados que criam obstáculos indesejáveis. Com o tempo, alguns desses picos e obstáculos adquirem um status icônico. São esses os grandes problemas.

O QUE TORNA um grande problema matemático grandioso? Profundidade intelectual, combinada com simplicidade e elegância. Mais: precisa ser *difícil.* Qualquer um pode escalar um morrinho; o Everest é algo inteiramente diferente. Um grande problema é geralmente simples de formular, embora os termos exigidos possam ser elementares ou altamente técnicos. As formulações do último teorema de Fermat e do problema das quatro cores podem fazer sentido imediato para qualquer pessoa familiarizada com matemática escolar. Em contraste, é impossível sequer formular a conjectura de Hodge ou a hipótese da diferença de massas sem invocar conceitos profundos nas fronteiras da pesquisa – a última, afinal, provém da teoria quântica de campo. Contudo, para aqueles que são versados em tais áreas, a formulação da questão envolvida é simples e natural. Não abrange páginas e páginas de texto denso e impenetrável. Entre os dois extremos, estão os problemas que requerem algo no nível de matemática de graduação, se você quiser entendê-los nos mínimos detalhes. Uma impressão mais geral sobre o que é essencial no problema – de onde veio, por que é importante, o que se poderia fazer se tivéssemos a solução – geralmente é acessível para qualquer pessoa interessada, e é isso que tentarei oferecer. Admito que, sob esse aspecto, a conjectura de Hodge é uma noz com a casca bem dura de ser quebrada. No entanto, é um dos sete problemas matemáticos do milênio do Instituto Clay, com um prêmio de 1 milhão de dólares, e realmente precisa ser incluído.

Grandes problemas são criativos: ajudam a dar à luz uma nova matemática. Em 1900, David Hilbert proferiu uma palestra no Congresso Internacional de Matemáticos em Paris, na qual listou 23 dos mais importantes problemas em matemática. Não incluiu o último teorema de Fermat, mas mencionou-o em sua introdução. Quando um matemático notável faz uma lista do que julga serem alguns dos grandes problemas, outros matemáticos prestam atenção. Os problemas não estariam na lista se não fossem importantes, e difíceis. É natural levantar-se ante os desafios, e tentar resolvê-los. Desde então, solucionar um dos problemas de Hilbert tem sido uma boa maneira de se receber condecorações matemáticas. Muitos desses problemas são técnicos demais para serem aqui incluídos, outros são esboços com final em aberto em vez de problemas específicos, e vários surgem mais tarde por si sós. Mas merecem ser mencionados, então coloquei um breve resumo nas notas.[7]

É isso que faz com que um grande problema matemático seja grandioso. Raras vezes, o que o torna problemático é decidir qual deve ser a resposta. Para praticamente todos os grandes problemas, os matemáticos têm uma ideia muito clara de qual deveria ser a resposta – ou tinham, se agora a solução já é conhecida. Com efeito, a formulação do problema muitas vezes inclui a resposta esperada. Qualquer coisa descrita como conjectura é algo desse tipo: um palpite plausível, baseado em uma variedade de evidências. A maioria das conjecturas bem estudadas acaba revelando-se correta, embora não todas. Termos mais antigos como "hipótese" carregam o mesmo significado, e no caso de Fermat a palavra "teorema" é (mais precisamente, era) um abuso – um teorema requer uma prova, mas era exatamente isso que faltava quando Wiles entrou em cena.

Uma prova, na verdade, é a exigência que torna problemáticos os grandes problemas. Qualquer indivíduo razoavelmente competente pode fazer alguns cálculos, identificar um padrão aparente e destilar sua essência numa formulação consistente. Os matemáticos exigem mais evidência do que isso: insistem em uma prova completa, com lógica impecável. Ou, se a resposta revelar-se negativa, uma refutação. Não é realmente possível, na empreitada matemática, contemplar o encanto sedutor de um grande problema sem apreciar o papel vital da prova. Qualquer pessoa pode dar um palpite. O difícil é provar que ele está certo. Ou errado.

O conceito de prova matemática tem mudado do decorrer da história, com as exigências lógicas tornando-se geralmente mais rigorosas. Já houve

muitas discussões filosóficas eruditas a respeito da natureza da prova, que levantaram algumas questões importantes. Definições lógicas precisas de "prova" têm sido propostas e implantadas. Aquela que ensinamos aos alunos de graduação é a de que uma prova tem início com uma coleção de premissas explícitas chamadas axiomas. Estes são, por assim dizer, as regras do jogo. Outros axiomas são possíveis, mas levam a jogos diferentes. Foi Euclides, o antigo geômetra grego, quem introduziu na matemática essa abordagem, válida até hoje. Estando de acordo quanto aos axiomas, a prova de alguma afirmação é uma série de passos, cada um sendo consequência lógica ou dos axiomas, ou de afirmações provadas anteriormente, ou de ambos. Na realidade, o matemático está explorando um labirinto lógico, cujos entroncamentos são afirmações e cujos corredores são deduções válidas. Uma prova é o caminho pelo labirinto, começando pelos axiomas. E o que ela prova é a afirmação no fim do labirinto.

Contudo, esse conceito límpido de prova não é a história toda. Não é sequer uma parte importante dela. É como dizer que uma sinfonia é uma sequência de notas musicais, sujeita a regras de harmonia. Essa definição não leva em conta nada da criatividade. Não nos diz como encontrar provas, nem mesmo como validar as provas de outras pessoas. Não nos diz que pontos no labirinto são significativos. Não nos diz quais trajetos são elegantes e quais são feios, quais são importantes e quais irrelevantes. É uma descrição formal, mecânica de um processo que possui muitos outros aspectos, sobretudo uma dimensão humana. Provas são descobertas por gente, e a pesquisa em matemática não é simplesmente uma questão de lógica passo a passo.

Considerar literalmente a definição formal de prova pode levar a provas ilegíveis, pelo fato de que a maior parte do tempo é gasta colocando os pingos nos *is* lógicos em circunstâncias em que o resultado já está diante de nós. Assim, matemáticos com prática vão direto ao assunto, deixando de fora tudo que é óbvio ou rotina. Eles deixam claro que há uma lacuna, lançando mão de frases feitas do tipo "é fácil verificar que" ou "cálculos de rotina mostram". O que eles não fazem, pelo menos não de maneira consciente, é resvalar numa dificuldade lógica e tentar fingir que ela não existe. Na verdade, um matemático competente sai do seu caminho para apontar exatamente aquelas partes do argumento que são logicamente frágeis, e dedica a maior parte do tempo a explicar como torná-las suficientemente consistentes. O desfecho é que a prova, na prática, é uma história matemática com seu próprio fluxo narrativo. Tem começo, meio e

fim. E, muitas vezes, tem tramas secundárias, ramificando-se do enredo principal, cada uma com suas próprias resoluções. O matemático britânico Christopher Zeeman comentou certa vez que um teorema é um ponto de repouso intelectual. Você pode parar, recuperar o fôlego e sentir que chegou a algum local definido. A trama secundária fica sendo uma ponta solta na história principal. Provas assemelham-se a narrativas também sob outros aspectos: com frequência têm um ou mais personagens centrais – ideias, em vez de pessoas, é claro – cujas complexas interações levam à revelação final.

Conforme indica a definição do curso de graduação, uma prova tem início com algumas premissas claramente formuladas, deduz consequências lógicas de maneira coerente e estruturada e termina com aquilo que se deseja provar. Mas uma prova não é apenas uma lista de deduções, e a lógica não é o único critério. Uma prova é uma história contada para e dissecada por pessoas que passaram muito tempo de sua vida aprendendo a como ler tais histórias e a achar erros ou inconsistências: pessoas cuja meta principal é provar que o narrador está *errado*, e que possuem a sinistra aptidão de identificar fraquezas e martelá-las até que desabem e se desfaçam numa nuvem de pó. Se algum matemático alegar que solucionou um problema importante, seja um dos grandes ou algo valioso porém menos badalado, o reflexo profissional não é gritar "oba!" e estourar uma garrafa de champanhe, e sim derrubar a resolução.

Isso pode soar negativo, mas a prova é a única ferramenta confiável de que os matemáticos dispõem para assegurar que aquilo que dizem está correto. Antecipando esse tipo de reação, os pesquisadores dedicam boa parte dos seus esforços tentando derrubar suas próprias ideias e provas. Dessa maneira, fica menos constrangedor. Quando a história consegue sobreviver a esse tipo de avaliação crítica, o consenso logo muda para concordar que está correta, e a essa altura seu inventor recebe os merecidos elogios, crédito e recompensa. Em todo caso, é desse modo que geralmente funciona, embora nem sempre possa parecer assim para os envolvidos. Se você está perto da ação, sua imagem do que está acontecendo pode ser diferente da imagem de um observador mais desligado.

COMO É QUE os matemáticos resolvem problemas? Até hoje houve poucos

estudos científicos a respeito dessa questão. A moderna pesquisa educacional, baseada em ciência cognitiva, é focada em grande parte na educação até o ensino médio. Alguns estudos abordam o ensino de matemática no nível de graduação universitária, mas são relativamente poucos. Há diferenças significativas entre aprender e ensinar a matemática existente e criar matemática nova. Muitos de nós sabemos tocar um instrumento musical, mas são poucos os que são capazes de compor um concerto ou mesmo escrever uma canção pop.

Quando se trata de criatividade em níveis mais elevados, muito do que sabemos – ou pensamos saber – vem da introspecção. Pedimos aos matemáticos para explicar seus processos de pensamento, e buscamos princípios gerais. Uma das primeiras tentativas sérias de descobrir como os matemáticos pensam foi o livro de Jacques Hadamard, *A psicologia da invenção na matemática*, publicado pela primeira vez em 1945.[8] Hadamard entrevistou proeminentes matemáticos e cientistas de sua época pedindo-lhes que descrevessem como pensavam quando trabalhavam em problemas difíceis. O que emergiu, com muita intensidade, foi o papel vital daquilo que, por falta de um termo melhor, precisa ser descrito como intuição. Algum elemento da mente subconsciente guiava seus pensamentos. Suas percepções mais criativas não surgiam por meio de uma lógica passo a passo, mas em saltos súbitos, descontrolados.

Uma das descrições mais detalhadas dessa abordagem aparentemente ilógica para questões lógicas foi dada pelo matemático francês Henri Poincaré, uma das figuras mais importantes do fim do século XIX e começo do século XX. Ele passeou através da maior parte da matemática, fundando diversas áreas e mudando radicalmente muitas outras. Desempenha papel de destaque em vários capítulos que virão. Também escreveu livros de ciência popular, e sua vasta gama de experiências pode tê-lo ajudado a adquirir uma compreensão mais profunda de seus próprios processos de pensamento. Em todo caso, Poincaré foi categórico em afirmar que a lógica consciente era apenas parte do processo criativo. Sim, havia momentos em que ela era indispensável: decidir qual era efetivamente o problema, verificar de maneira sistemática a resposta. Mas nesse entremeio, Poincaré sentia que seu cérebro muitas vezes estava trabalhando num problema sem lhe contar, de um modo que ele simplesmente não conseguia imaginar.

Sua delineação do processo criativo distinguia três estágios principais: preparação, incubação e iluminação. A preparação consiste em esforços lógicos conscientes para especificar o problema, torná-lo preciso e atacá-lo

pelos métodos convencionais. Poincaré considerava esse estágio essencial: ele dá a partida no subconsciente e provê a matéria-prima para ele trabalhar. A incubação tem lugar no momento em que você para de pensar sobre o problema e começa a se dedicar a outra coisa. O subconsciente agora passa a combinar ideias entre si, muitas vezes ideias em estado bruto, até começar a surgir uma luz. Com sorte, isso conduz à iluminação: o subconsciente lhe dá um tapinha no ombro e a lâmpada proverbial se apaga em sua mente.

Esse tipo de criatividade é como andar na corda bamba. Por um lado, não se resolve um problema difícil a menos que você se familiarize com a área ao qual ele pertence – junto com muitas outras áreas, que podem ou não estar relacionadas, mas, caso estejam, você vai precisar delas. Por outro lado, se tudo que você fizer ficar aprisionado nos modos padronizados de pensar que os outros já tentaram infrutiferamente então você acabará encalhado num sulco mental sem descobrir nada de novo. Assim, o truque é saber muito, integrar conscientemente o que sabe e botar seu cérebro em funcionamento durante semanas … e aí deixar a questão de lado. A parte intuitiva da sua mente começa então a trabalhar; esfrega as ideias umas contra as outras para ver de onde sai alguma faísca e o avisa quando encontrou algo. Isso pode acontecer a qualquer momento: Poincaré descobriu de repente como resolver um problema que o vinha incomodando por meses, simplesmente quando descia do ônibus. Srinivasa Ramanujan, um matemático indiano autodidata com talento para fórmulas notáveis, tinha com frequência suas ideias em sonhos. É famoso como Arquimedes descobriu a maneira de se testar um metal para ver se era ouro ao tomar banho.

Poincaré deu-se ao trabalho de ressaltar que sem o período inicial de preparação, o progresso é improvável. O subconsciente, insistiu ele, necessita receber farto material para pensar, caso contrário a fortuita combinação de ideias que eventualmente levará a uma solução não pode se formar. A transpiração gera a inspiração. Ele também deve ter sabido – pois todo matemático criativo sabe – que é raro que esse processo em três estágios ocorra apenas uma vez. Solucionar um problema geralmente requer mais do que uma grande sacada. O estágio de incubação para uma ideia pode ser interrompido por um processo subsidiário de preparação, incubação e iluminação de algo necessário para fazer a primeira ideia funcionar. A solução de qualquer problema digno de esforço, seja grande ou não, envolve tipicamente muitas dessas sequências, aninhadas uma dentro

da outra como os intricados fractais de Benoît Mandelbrot. Você resolve um problema decompondo-o em subproblemas. E se convence de que, se resolver esses subproblemas, poderá juntar os resultados para resolver a coisa inteira. Então você trabalha nos subproblemas. Às vezes resolve um; outras vezes fracassa, e cabe então repensá-lo. Às vezes um subproblema em si decompõe-se em mais pedaços. Pode dar muito trabalho simplesmente seguir o plano traçado.

Descrevi o funcionamento do subconsciente como "intuição". Esta é uma daquelas palavras sedutoras como "instinto", que é usada de maneira ampla mesmo sendo destituída de qualquer sentido real. É um nome para algo cuja presença reconhecemos, mas não entendemos. A intuição matemática é a capacidade mental de sentir forma e estrutura, detectar padrões que não conseguimos perceber conscientemente. A intuição carece da clareza cristalina da lógica consciente, mas compensa essa falta chamando a atenção para coisas que nunca teríamos considerado de modo consciente. Os neurocientistas mal estão começando a compreender como o cérebro executa tarefas muito mais simples. Mas seja lá como a intuição funcione, ela deve ser uma consequência da estrutura do cérebro e da maneira como ele interage com o mundo exterior.

MUITAS VEZES a contribuição-chave da intuição é conscientizar-nos dos pontos fracos em um problema, lugares onde ele possa estar vulnerável ao ataque. Um prova matemática é como uma batalha ou, se você preferir, uma metáfora menos bélica, um jogo de xadrez. Uma vez identificado um ponto fraco, o domínio técnico que o matemático tem do maquinário da matemática pode ser colocado em ação para explorá-lo. Como Arquimedes, que queria um ponto de apoio firme para poder mover a Terra, o pesquisador matemático necessita ter algum meio de alavancar o problema. Uma ideia-chave pode abri-lo, tornando-o vulnerável aos métodos-padrão. Depois disso, é só uma questão de técnica.

Meu exemplo favorito desse tipo de alavancagem é um quebra-cabeça que não possui significação matemática intrínseca, mas encerra uma importante mensagem. Suponha que você tenha um tabuleiro de xadrez, com 64 casas, e um suprimento de dominós do tamanho exato para cobrir dois quadrados adjacentes do tabuleiro. Então é fácil cobrir o tabuleiro inteiro com 32 dominós. Mas agora suponha que dois cantos diagonalmente opostos do tabuleiro sejam removidos, como na Figura 1.

Será que as 62 casas restantes podem ser cobertas usando 31 dominós? Se você experimentar, parece que nada dá certo. Por outro lado, é difícil enxergar alguma razão óbvia para que a tarefa seja impossível. Até você perceber que, de qualquer maneira que as pedras sejam arranjadas, cada uma delas precisa cobrir um quadrado preto e um branco. Aí está sua alavanca; tudo que você precisa fazer agora é manejá-la. Ela implica que qualquer região coberta por dominós contenha o mesmo número de quadrados pretos e brancos. Mas quadrados diagonalmente opostos têm a mesma cor, de modo que ao remover dois deles (no caso do exemplo, dois brancos) obtemos uma forma com dois quadrados pretos a mais que brancos. Então essa forma não pode ser coberta. A observação sobre a combinação de cores cobertas por *qualquer* dominó é o ponto fraco do quebra-cabeça. Ela fornece um lugar para você apoiar sua alavanca lógica e empurrar. Se você fosse um barão medieval atacando um castelo, este seria o ponto fraco da muralha – o lugar onde você deveria concentrar o poder de fogo de suas catapultas, ou cavar um túnel para miná-la.

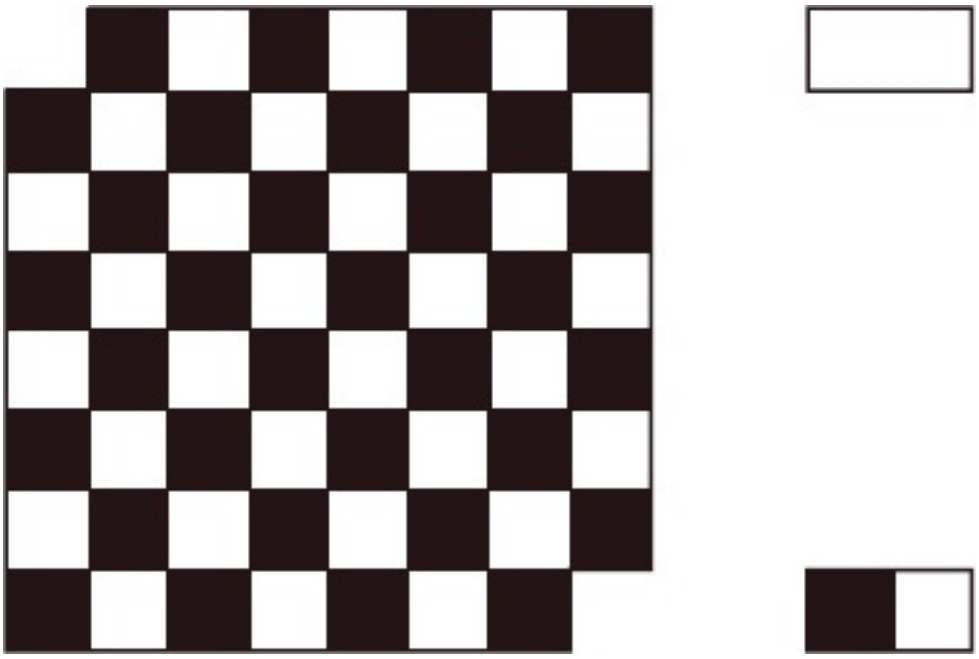


FIGURA 1 Você consegue cobrir com dominós o tabuleiro recortado, cada um cobrindo dois quadrados (à dir., no alto)? Se você colorir o dominó (à dir., embaixo) e contar quantos quadrados pretos e brancos há, a resposta fica clara.

A pesquisa matemática difere de uma batalha em um aspecto importante. Qualquer território que você ocupe, permanece seu para sempre. Você pode decidir concentrar seus esforços em alguma outra coisa, mas uma vez que um teorema é provado, ele jamais volta a desaparecer. É

assim que os matemáticos fazem progresso em um problema, mesmo quando fracassam em resolvê-lo. Eles estabelecem um fato novo, que fica então à disposição de qualquer um para ser usado, em qualquer contexto que seja. Muitas vezes, a plataforma de lançamento de um novo ataque a um problema antigo surge a partir de alguma preciosidade anteriormente despercebida, meio soterrada numa pilha de fatos diversificados. E essa é uma das razões para a matemática nova ser importante por si só, mesmo que seus usos não sejam imediatamente visíveis. É mais um pedaço de território ocupado, mais uma arma no arsenal. Sua vez pode ainda chegar – porém, com certeza, não chegará caso seja considerada "inútil" e esquecida, ou jamais permitida de vir a existir pelo fato de que não se consegue ver para o que ela *serve*.

---

[a] British Academy of Film and Television Arts (Academia Britânica de Artes do Cinema e Televisão), que premia anualmente programas que se destacam nos meios audiovisuais. (N.T.)

2. Território dos primos

A conjectura de Goldbach

ALGUNS DOS GRANDES problemas matemáticos manifestam-se bem cedo na nossa educação matemática, embora possamos nem notar. Logo depois que aprendemos a multiplicação, deparamos com o conceito de número primo. Alguns números podem ser obtidos pela multiplicação de dois números menores, por exemplo, $6 = 2 \times 3$. Outros, tais como 5, não podem ser decompostos dessa maneira; o máximo que podemos fazer é $5 = 1 \times 5$, que não envolve dois números *menores*. Números que podem ser divididos em dois menores são ditos compostos; os que não podem são primos. Números primos parecem coisas tão simples. Logo que você aprende a multiplicar dois números inteiros, é capaz de entender o que é um número primo. Os primos são os blocos construtivos básicos dos números inteiros, e estão presentes em toda a matemática. E são também profundamente misteriosos, parecendo estar espalhados quase ao acaso. Não há dúvida quanto a isso: eles são um enigma. Talvez seja consequência de sua definição – não tanto o que são quanto o que não são. Por outro lado, são fundamentais para a matemática, de modo que não podemos simplesmente erguer os braços horrorizados e desistir. Precisamos chegar a um acordo com os números primos, e desvendar seus segredos mais íntimos.

Algumas características são óbvias. Com exceção do menor primo, 2, todos os outros são ímpares. Com exceção do 3, a soma de seus algarismos não pode ser múltipla de 3. Com exceção do 5, não podem terminar em 5. Exceto essas regras, e algumas outras mais sutis, não se pode olhar para um número e identificar de imediato se ele é primo. Existem, sim, fórmulas para os primos, mas em grande medida elas são tapeações: não fornecem informação nova útil sobre os primos; são apenas meios astutos de codificar a definição de "primo" em uma fórmula. Os primos são como as pessoas: são indivíduos, e não se adaptam a regras padronizadas.

Com o correr dos milênios, os matemáticos foram aos poucos aumentando sua compreensão dos números primos, e de tempos em tempos algum grande problema sobre eles é resolvido. No entanto, muitas perguntas ainda permanecem sem resposta. Algumas são básicas e fáceis

de formular; outras são mais esotéricas. Este capítulo discute o que sabemos e o que não sabemos a respeito desses números irritantes, todavia fundamentais. O capítulo começa estabelecendo alguns dos conceitos básicos, em particular a decomposição em fatores primos, ou fatoração – como exprimir um dado número multiplicando números primos entre si. Mesmo esse processo familiar conduz a águas profundas assim que começamos a solicitar métodos efetivos para achar os fatores primos de um número. Uma surpresa é que parece ser relativamente fácil testar um número para determinar se ele é primo, mas se for composto, achar seus fatores primos é geralmente muito mais difícil.

Tendo estabelecido a base, passamos para o mais famoso problema não resolvido sobre os primos, a conjectura de Goldbach, que conta 250 anos de existência. Progressos recentes sobre essa questão têm sido dramáticos, mas ainda não decisivos. Alguns outros problemas fornecem uma breve amostra do que ainda está por ser descoberto a respeito dessa rica, mas insubmissa, área da matemática.

OS NÚMEROS PRIMOS e a fatoração nos são familiares da aritmética da escola, mas a maioria das características importantes dos primos raramente é ensinada nesse nível, e quase nada é provado. Existem sólidas razões para isso: as provas, mesmo de propriedades aparentemente óbvias, são surpreendentemente difíceis. Em vez disso, os alunos aprendem alguns métodos simples para trabalhar com primos. Como resultado, a nossa experiência inicial com esses números é um tanto enganosa.

Os antigos gregos conheciam algumas das propriedades básicas dos primos, e sabiam como prová-las. Primos e fatores são o tópico principal do Livro VII dos *Elementos* de Euclides, o grande clássico da geometria. Esse livro específico contém uma apresentação geométrica da divisão e multiplicação em aritmética. Os gregos preferiam trabalhar com comprimentos de linhas, em vez dos números como tais, mas é fácil reformular seus resultados na linguagem dos números. Euclides tem o cuidado de provar afirmações que podem parecer óbvias: por exemplo, a Proposição 16 do Livro VII prova que quando dois números são multiplicados entre si, o resultado é independente da ordem em que são tomados. Isto é, $ab = ba$, uma das leis básicas da álgebra.

Em aritmética escolar, os fatores primos são usados para achar o máximo divisor comum (ou maior fator comum) de dois números. Por

exemplo, para encontrar o máximo divisor comum de 135 e 630, nós decompomos ambos os números em fatores primos:

$135 = 3^3 \times 5 \qquad 630 = 2 \times 3^2 \times 5 \times 7$

Então, para cada primo da decomposição, pegamos a maior potência que ocorre em ambas as fatorações, obtendo $3^2 \times 5$. Multiplicando, obtemos 45: este é o máximo divisor comum. Esse procedimento dá a impressão de que a decomposição em fatores primos é necessária para achar máximos divisores comuns. Na verdade, a relação lógica faz o trajeto inverso. A Proposição 2 do Livro VII dos *Elementos* apresenta um método para achar o máximo divisor comum de dois números inteiros sem fatorá-los. Ele funciona subtraindo-se repetidamente o número menor do maior, e aí aplicando a mesma operação ao resto resultante da primeira subtração e o número menor, continuando o processo até não haver mais resto. Para 135 e 630, um exemplo típico usando números pequenos, o processo se desenrola da seguinte maneira. Vamos subtrair repetidamente 135, primeiro de 630, depois dos restos resultantes:

$630 - 135 = 495$
$495 - 135 = 360$
$360 - 135 = 225$
$225 - 135 = 90$

Como 90 é menor que 135, trocamos os dois números de lugar:

$135 - 90 = 45$

Como 45 é menor que 90, trocamos os dois de lugar:

$90 - 45 = 45$
$45 - 45 = 0$

Portanto, o máximo divisor comum de 135 e 630 é 45.

Esse procedimento funciona porque em cada etapa ele substitui o par de números original por um par mais simples (um dos números é menor) que tem o mesmo máximo divisor comum. Um dos números acaba

dividindo o outro exatamente, e aí nós paramos. O termo atual para um método computacional explícito que tenha a garantia de achar a resposta para um dado problema é “algoritmo”. Então, o procedimento de Euclides é agora chamado de algoritmo euclidiano. Ele é logicamente anterior à decomposição em fatores primos. De fato, Euclides usa seu algoritmo para provar propriedades básicas dos fatores primos, e o mesmo fazem hoje os cursos universitários de matemática.

A Proposição 30 de Euclides é fundamental para o conjunto da empreitada. Em termos modernos, afirma que se um número primo divide o produto de dois números – o resultado da multiplicação de um pelo outro – então ele deve dividir um dos dois. A Proposição 32 afirma que um número ou é primo ou possui um fator primo. Juntando as duas, é fácil deduzir que todo número é um produto de fatores primos, e que essa expressão é única, e independe da ordem em que os fatores sejam escritos. Por exemplo:

$$60 = 2 \times 2 \times 3 \times 5 = 2 \times 3 \times 2 \times 5 = 5 \times 3 \times 2 \times 2$$

e assim por diante, mas a única maneira de obter 60 é rearranjar a primeira fatoração. Não há fatoração, por exemplo, que tenha alguma coisa do tipo 60 = 7 × *algo*. A existência da fatoração vem da Proposição 32. Se o número é primo, pare. Se não, ache um fator primo, divida para obter um número menor e repita. A singularidade vem da Proposição 30. Por exemplo, se houvesse uma fatoração 60 = 7 × *algo*, então 7 teria de dividir um dos números 2, 3 ou 5, o que não ocorre.

A essa altura preciso esclarecer um pequeno, mas importante, detalhe: o status excepcional do número 1. Segundo a definição até agora, 1 é claramente primo: se tentarmos decompô-lo, o melhor que podemos fazer é 1 = 1 × 1, que não envolve números menores. No entanto, mais tarde, essa interpretação causa problemas na teoria, então, nos últimos um ou dois séculos, os matemáticos acrescentaram uma restrição adicional. O número 1 é tão especial que não deve ser considerado nem primo nem composto. Em vez disso, é uma terceira espécie, uma unidade. Um dos motivos para tratar o 1 como caso especial, e não como um primo genuíno, é o fato de que se o chamarmos de primo, a singularidade falha. De fato, 1 × 1 = 1 já exibe a falha, e 1 × 1 × 1 × 1 × 1 × 1 × 1 × 1 = 1 a esfrega na nossa cara. Poderíamos modificar a singularidade e dizer “única, exceto números 1 adicionais”, mas este é simplesmente outro modo de admitir que 1 é especial.

Muito mais tarde, na Proposição 20 do Livro IX, Euclides prova outro fato-chave: “Os números primos são mais do que qualquer imensidão de números primos.” Ou seja, a quantidade de primos é infinita. Trata-se de um magnífico teorema com uma prova sagaz, mas abriu uma caixa de Pandora. Se os primos continuam para sempre, e todavia parecem não ter padrão, como podemos descrever sua aparência?

PRECISAMOS ENCARAR essa pergunta porque não podemos ignorar os números primos. Eles são traços essenciais da paisagem matemática. São especialmente comuns, e úteis, na teoria dos números. Essa área da matemática estuda as propriedades dos números inteiros. Pode parecer um pouco elementar, mas na verdade a teoria dos números é uma das áreas mais profundas e mais difíceis da matemática. Mais adiante vamos ver evidências de sobra acerca dessa afirmação. Em 1801, Gauss, o principal teórico dos números de sua época – indiscutivelmente um dos principais matemáticos de todos os tempos, talvez mesmo o maior de todos – escreveu um livro-texto avançado sobre teoria dos números, o *Disquisitiones Arithmeticae* (Investigações em aritmética). Entre tópicos de alto nível, ele ressaltou que não devemos perder de vista duas questões muito básicas: “O problema de distinguir os números primos dos números compostos e o de resolver os últimos em seus fatores primos é conhecido como um dos mais importantes e mais úteis em aritmética.”

Na escola, é comum nos ensinarem um modo de encontrar os fatores primos de um número: experimentar todos os fatores possíveis, um de cada vez, até achar algo apropriado. Se você não encontrou um fator até a hora em que atinge a raiz quadrada do número original – mais precisamente, o maior número inteiro que seja menor ou igual a essa raiz quadrada – então o número é primo. Se você achar um fator, divida por ele e repita a operação. É mais eficiente tentar apenas fatores primos, o que requer ter uma lista deles. Você cessa na raiz quadrada porque o menor fator de qualquer número composto não é maior que sua raiz quadrada. Contudo, esse procedimento é irremediavelmente ineficiente quando os números tornam-se grandes. Por exemplo, se o número é

1.080.813.321.843.836.712.253

então sua decomposição em fatores primos é

13.929.010.429 × 77.594.408.257

e você teria de experimentar os primeiros 624.401.249 primos, um de cada vez, para achar o menor dos dois fatores. É claro que com um computador isso é bastante fácil, mas se começarmos com um número de cem dígitos que por acaso seja o produto de dois números de cinquenta dígitos, e empregarmos uma busca sistemática através de primos sucessivos, o universo acaba antes que o computador ache a resposta.

Na verdade, os computadores de hoje geralmente podem fatorar números de cem dígitos. O meu leva menos do que um segundo para achar os fatores primos de $10^{99} + 1$, que é algo como 1.000 ... 001 com 98 zeros. É um produto de treze primos (um deles ocorre duas vezes), dos quais o menor é 7 e o maior é

141.122.524.877.886.182.282.233.539.317.796.144.938.305.111.168.71

Mas se eu mandar o computador fatorar $10^{199} + 1$, com duzentos dígitos, ele fica dando voltas para sempre sem chegar a lugar algum. Mesmo assim, o cálculo com cem dígitos é impressionante. Qual é o segredo? Achar métodos mais eficientes do que tentar todos os fatores primos potenciais um de cada vez.

Sabemos agora muito mais do que Gauss sobre o seu primeiro problema (testar os primos) e muito menos do que gostaríamos a respeito do segundo (fatoração). A sabedoria convencional é que testar primos é muito mais simples do que a fatoração. Isso geralmente surpreende aos não matemáticos, que aprenderam na escola a testar se um número é primo pelo mesmo método usado para fatorar: experimentar todos os divisores possíveis. Acontece que há modos habilidosos de provar que um número é primo sem fazer isso. E que também nos permitem provar que um número é composto, sem achar nenhum dos seus fatores. Basta mostrar que ele não passa por um teste de primos.

O grande tataravô de todos os testes de verificação de primos é o teorema de Fermat, que não deve ser confundido com seu celebrado último teorema (Capítulo 7). Esse teorema baseia-se em aritmética modular, às vezes conhecida como “aritmética do relógio” porque os números dão voltas como em um relógio. Pegue um número – para um relógio analógico de doze horas é o número 12 – e chame-o de módulo. Em qualquer cálculo aritmético com números inteiros, você agora se permite substituir qualquer múltiplo de 12 por zero. Por exemplo, $5 \times 5 = 25$, mas 24 é duas vezes 12, então subtraindo 24 obtemos $5 \times 5 = 1$ em módulo 12. A

aritmética modular é muito bonita, porque quase todas as regras habituais da aritmética ainda funcionam. A principal diferença é que não se pode sempre dividir um número por outro, mesmo quando não é zero. A aritmética modular também é útil, porque fornece um modo bem-organizado de lidar com questões de divisibilidade: que números são divisíveis pelo módulo escolhido, e qual é o resto quando não são? Gauss introduziu a aritmética modular no *Disquisitiones Arithmeticae*, e hoje ela é amplamente usada em ciência de computação, física e engenharia, bem como em matemática.

O teorema de Fermat afirma que se escolhermos um módulo primo p, e pegarmos qualquer número a que não seja múltiplo de p, então a potência $(p - 1)$ de a é igual a 1 em aritmética de módulo p. Suponhamos, por exemplo, que $p = 17$ e $a = 3$. Então o teorema prediz que quando dividimos 3^{16} por 17, o resto é 1. Como verificação,

$$3^{16} = 43.046.721 = 2.532.160 \times 17 + 1$$

Ninguém em pleno poder de suas faculdades mentais gostaria de fazer contas desse tipo com, digamos, primos de cem dígitos. Felizmente, existe um modo astucioso e rápido para realizar esse tipo de operação. O ponto é que, se a resposta não for igual a 1, então o módulo com que começamos é um número composto. Assim, o teorema de Fermat constitui a base para um teste eficiente que fornece a condição necessária para um número ser primo.

Infelizmente, o teste não é suficiente. Muitos números compostos, conhecidos como números de Carmichael, passam no teste. O menor é 561, e em 2003 Red Alford, Andrew Granville e Carl Pomerance provaram, para espanto geral, que eles são infinitos. A reação foi pelo fato de eles terem achado uma prova; o resultado em si não foi tão surpreendente. Na verdade, mostraram que existem pelo menos $x^{2/7}$ números de Carmichael a menos ou em quantidade igual a x se x for grande o bastante.

Entretanto, variantes mais sofisticadas do teorema de Fermat podem ser transformadas em testes genuínos de primalidade, tais como a publicada em 1976 por Gary Miller. Infelizmente, a prova da validade do teste de Miller depende de um grande problema não resolvido, a hipótese generalizada de Riemann (Capítulo 9). Em 1980, Michael Rabin transformou o teste de Miller em um teste probabilístico, que ocasionalmente podia dar uma resposta errada. As exceções, se existirem,

são muito raras, mas não podem ser totalmente excluídas. O teste determinístico (isto é, com garantia de ser correto) mais eficiente até o momento é o de Adleman-Pomerance-Rumely, batizado em homenagem a Leonard Adleman, Carl Pomerance e Robert Rumely. Ele utiliza ideias da teoria dos números que são mais sofisticadas do que o teorema de Fermat, mas num espírito similar.

AINDA ME RECORDO com clareza de uma carta de um amador esperançoso, que propôs uma variante da divisão por tentativas. Experimente todos os divisores possíveis, mas comece pela raiz quadrada e trabalhe *de cima para baixo*. Esse método às vezes chega à resposta mais depressa do que fazer as coisas na ordem usual, mas à medida que os números vão ficando maiores, acabamos deparando com a mesma dificuldade que o método habitual. Se você fizer a tentativa pelo meu exemplo acima, o número de 22 dígitos 1.080.813.321.843.836.712.253, então a raiz quadrada é mais ou menos 32.875.725.419. Você precisa tentar 794.582.971 divisores primos antes de achar um que dê certo. Isso é *pior* do que procurar pela forma usual.

Em 1956 o famoso lógico Kurt Gödel, escrevendo para John von Neumann, fez eco ao apelo de Gauss. Perguntou se a divisão por tentativas podia ser melhorada, e, em caso positivo, em quanto. Von Neumann não se dedicou à questão, mas com os anos outros responderam a Gödel descobrindo métodos práticos de achar primos com até cem dígitos, às vezes mais. Esses métodos, dos quais o mais famoso é chamado crivo quadrático, são conhecidos desde cerca de 1980. No entanto, quase todos são ou probabilísticos ou ineficientes no seguinte sentido.

Como é que o tempo de processamento de um algoritmo de computador cresce à medida que aumenta a quantidade de dados de entrada – o *input*? Para testar primos, o tamanho do *input* não é o número em si, mas quantos dígitos ele tem. A distinção básica em tais questões é entre duas classes de algoritmos chamados P e não-P. Se o tempo de processamento aumenta como alguma potência fixa do tamanho do *input*, então o algoritmo é classe P; do contrário, é não-P. Grosso modo, algoritmos classe P são úteis, ao passo que algoritmos não-P não são práticos, mas entre eles há algo como "uma terra de ninguém", onde entram em jogo outras considerações. Aqui P significa "tempo polinomial", uma maneira enfeitada de falar de potências, e voltaremos ao tópico de algoritmos eficientes no Capítulo 11.

Pelo padrão classe P, a divisão por tentativas tem um desempenho

péssimo. Em sala de aula, onde os números que ocorrem têm dois ou três dígitos, isso funciona, mas é completamente inviável com números de cem dígitos. A divisão por tentativas acha-se, com certeza, na classe não-P. Na verdade, o tempo de processamento é de aproximadamente $10^{n/2}$ para um número de n dígitos, o que cresce mais rápido do que qualquer potência fixa de n. Esse tipo de crescimento, chamado exponencial, é *realmente* terreno computacional nebuloso, ruim.

Até a década de 1980, todos os algoritmos conhecidos para testar números primos, excluindo os probabilísticos ou aqueles cuja validade não foi provada, tiveram taxa de crescimento exponencial. Contudo, em 1983, foi descoberto um algoritmo naquela lacuna adjacente ao território P, chamada "terra de ninguém": o já mencionado teste de Adleman-Pomerance-Rumely. Uma versão melhorada, de Henri Cohen e Hendrik Lenstra, tem tempo de processamento n elevado à potência ln ln n, onde ln significa o logaritmo natural. Tecnicamente, ln ln n pode ser tão grande quanto se queira, de modo que esse algoritmo não é classe P. Mas isso não impede que seja prático: se n for um googolplex, 1 seguido de 10^{100} zeros, então ln ln n vale aproximadamente 230. Uma antiga piada diz: "Foi provado que ln ln n tende a infinito, mas ele nunca foi observado fazendo isso."

O primeiro teste de primalidade na classe P foi descoberto em 2002 por Manindra Agrawal e seus alunos Neeraj Kayal e Nitin Saxena, que na época eram estudantes de graduação. Conto alguns detalhes nas notas.[1] Eles provaram que seu algoritmo tinha um tempo de processamento proporcional a no máximo n^{12}, e isso foi rapidamente melhorado para $n^{7,5}$. No entanto, mesmo que seu algoritmo seja classe P, portanto classificado como "eficiente", suas vantagens não aparecem até o número n tornar-se de fato muito grande. Ele deve bater o teste Adleman-Pomerance-Rumely quando o número de dígitos n for em torno de $10^{1.000}$. Não existe espaço para encaixar um número desse tamanho na memória de um computador, ou, na verdade, no universo conhecido. Porém, agora que *sabemos* que existe um algoritmo classe P para testar números primos, começa a valer a pena procurar outros melhores. Lenstra e Pomerance reduziram a potência de 7,5 para 6. Se várias outras conjecturas sobre primos forem verdadeiras, então a potência poderá ser reduzida para 3, o que começa a parecer prático.

O aspecto mais empolgante do algoritmo Agrawal-Kayal-Saxena, porém, não é o resultado, mas o método. É um método simples – bem, pelo menos

para os matemáticos – e com novidades. A ideia subjacente é uma variante do teorema de Fermat, mas em vez de trabalhar com números, o grupo de Agrawal usou um polinômio, que é uma combinação de potências da variável x, tal como $5x^3 + 4x - 1$. Podemos somar, subtrair e multiplicar polinômios, e as leis algébricas habituais permanecem válidas. O Capítulo 3 explica polinômios em mais detalhes.

Essa é realmente uma ideia bacana: expandir o domínio do discurso e transportar o problema para um novo campo de pensamento. É uma dessas ideias que são tão simples que é preciso ser um gênio para enxergá-la. Ela evoluiu de um artigo de 1999 de autoria de Agrawal e de seu supervisor de doutorado, Somenath Biswas, fornecendo um teste probabilístico de primalidade baseado em um análogo ao teorema de Fermat no mundo dos polinômios. Agrawal estava convencido de que o elemento probabilístico podia ser removido. Em 2001 seus alunos surgiram com uma observação crucial, bastante técnica. Prosseguindo nesse caminho, o grupo foi levado às profundezas da teoria dos números, mas por fim tudo acabou se reduzindo a um único obstáculo, a existência de um primo p tal que $p - 1$ tenha um divisor primo suficientemente grande. Algumas poucas indagações e buscas na internet levaram a um teorema provado por Etienne Fouvry em 1985 usando métodos e técnicas profundos. Era exatamente o que precisavam para provar que seu algoritmo funcionava, e a peça final do quebra-cabeça encaixou-se precisamente no lugar.

NOS DIAS EM QUE a teoria dos números estava isolada em sua própria torrezinha de marfim, nada disso teria tido importância para o resto do mundo. Mas nos últimos vinte anos, os números primos tornaram-se importantes em criptografia, a ciência dos códigos secretos. Códigos não são importantes apenas para fins militares; empresas comerciais também têm seus segredos. Na era da internet, todos nós temos: não queremos que criminosos tenham acesso a nossas contas bancárias, números de cartões de crédito, ou, com o aumento dos roubos de identidade, ao nome do nosso gato. Mas a internet é um modo tão conveniente de pagar contas, fazer seguro de carros ou reservas para as férias, que temos de aceitar algum risco de nossa informação pessoal e confidencial cair nas mãos erradas.

Fabricantes de computadores e provedores de serviços na internet tentam reduzir esse risco viabilizando vários sistemas de encriptação. O envolvimento dos computadores mudou tanto a criptografia como a

criptoanálise, a sombria arte da quebra de códigos. Muitos códigos novos têm sido concebidos, e um dos mais famosos, inventado por Ron Rivest, Adi Shamir e Leonard Adleman em 1978, usa números primos. Números primos grandes, com cerca de cem dígitos. O sistema Rivest-Shamir-Adleman é empregado em muitos sistemas operacionais de computação, embutido nos principais protocolos de comunicação de internet segura, e amplamente usado por governos, corporações e universidades. Isso não quer dizer que todo novo resultado referente a números primos seja significativo para a segurança da sua conta bancária na internet, mas decididamente adiciona um elemento de excitação a qualquer descoberta que relacione primos com computação. O teste Agrawal-Kayal-Saxena é um desses casos. Matematicamente, é elegante e importante, mas não tem significação prática direta.

Ele lança, porém, uma luz nova e ligeiramente perturbadora sobre a questão geral da criptografia Rivest-Shamir-Adleman. Ainda não existe algoritmo classe P para resolver o segundo problema de Gauss, a fatoração. A maioria dos estudiosos pensa não haver nada desse tipo, mas já não têm a mesma certeza de antes. Uma vez que novas descobertas, como o teste Agrawal-Kayal-Saxena, podem estar ocultas espreitando, baseadas em ideias simples tais como versões polinomiais do teorema de Fermat, criptossistemas baseados em fatoração de números primos podem não ser tão seguros como ingenuamente imaginamos. Continue sem revelar o nome do seu gato na internet.

MESMO A MATEMÁTICA básica dos números primos conduz rapidamente a conceitos mais avançados. O mistério fica ainda mais profundo quando fazemos perguntas mais sutis. Euclides provou que os primos continuam para sempre, portanto não podemos apenas listá-los e pronto. Tampouco podemos dar uma fórmula algébrica simples para números primos sucessivos, da mesma maneira que x^2 especifica quadrados. (Existem, sim, fórmulas simples, mas elas "tapeiam" introduzindo os primos de maneira disfarçada e não nos apresentam nada de novo.[2]) Para apreender a natureza desses números erráticos, elusivos, podemos realizar experimentos, procurar pistas da estrutura e tentar provar que esses padrões aparentes persistem não importa quão grandes os números primos se tornem. Por exemplo, podemos perguntar como eles se distribuem entre todos os números naturais. Tabelas de primos sugerem com veemência que eles tendem a escassear à medida que vão

aumentando. A Tabela 1 mostra quantos deles existem nas várias faixas de mil números consecutivos.

FAIXA	QUANTIDADE DE PRIMOS
1 – 1.000	168
1.001 – 2.000	135
2.001 – 3.000	127
3.001 – 4.000	119
4.001 – 5.000	118
5.001 – 6.000	114
6.001 – 7.000	117
7.001 – 8.000	106
8.001 – 9.000	110
9.001 – 10.000	111

TABELA 1 A quantidade de primos em intervalos sucessivos de mil números.

Os números da segunda coluna, em sua maior parte, decrescem à medida que vamos descendo as linhas, embora às vezes haja breves períodos em que aumentam: por exemplo, 114 é seguido de 117. Esse é um sintoma da irregularidade dos primos, mas, apesar disso, há uma tendência geral de que se tornem mais raros à medida que seu tamanho aumenta. O motivo não é difícil de identificar: quanto maior o número, mais fatores potenciais ele tem. Primos precisam evitar todos esses fatores. É como pescar não primos com uma rede: quanto mais fina, menos primos escapam.

A "rede" tem até nome: a peneira de Eratóstenes. Eratóstenes de Cirene era um matemático da Grécia antiga que viveu por volta de 250 a.C. Era também atleta com interesse em poesia, geografia, astronomia e música. Foi dele a primeira estimativa razoável do tamanho da Terra, observando a posição do Sol ao meio-dia em dois locais diferentes, Alexandria e Siene – atualmente Assuã. Ao meio-dia, o Sol estava exatamente a pino em Siene, mas a sete graus da vertical em Alexandria. Considerando que esse ângulo é $^{1}/_{50}$ de um círculo, a circunferência da Terra devia ser cinquenta vezes maior que a distância de Alexandria a Siene. Eratóstenes não podia medir a distância diretamente, então perguntou a mercadores quanto tempo levava para fazer a viagem de camelo, e estimou a distância geralmente percorrida tipicamente em um dia. Ele deu um número explícito numa unidade conhecida como *estádio*, mas não sabemos qual o comprimento dessa

unidade. Os historiadores geralmente pensam que a estimativa de Eratóstenes era relativamente acurada.

Sua peneira é um algoritmo para encontrar todos os números primos eliminando sucessivamente todos os múltiplos já conhecidos como primos. A Figura 2 ilustra o método para os números até 102, dispostos de maneira a tornar o processo de eliminação fácil de acompanhar. Para ver o que acontece, sugiro que você mesmo construa o diagrama. Comece apenas com o reticulado, omitindo as linhas que riscam os números eliminando-os. Então adicione essas linhas uma por uma. Omita o 1 porque é a unidade. O número seguinte é 2, então é primo. Risque todos os múltiplos de 2: os que estão nas linhas horizontais começando por 4, 6 e 8. O número seguinte não riscado é 3, então é primo. Risque todos os múltiplos de 3: os que estão na linha horizontal começando por 6, já riscado, e 9. O número seguinte não riscado é 5, então é primo. Risque todos os múltiplos de 5: os que estão nas linhas diagonais subindo para a direita, começando por 10. O número seguinte não riscado é 7, então é primo. Risque todos os múltiplos de 7: os que estão nas linhas diagonais descendo para a direita, começando por 14. O número seguinte não riscado é 11, então é primo. O primeiro múltiplo de 11 que ainda não foi riscado por ter um divisor menor é 121, que está fora da figura, então pare. Os números restantes, sombreados, são os primos.

1	7	13	19	~~25~~	31	37	43	~~49~~	~~55~~	61	67	73	79	~~85~~	91	97
2	~~8~~	~~14~~	~~20~~	~~26~~	32	38	44	~~50~~	~~56~~	62	68	74	~~80~~	86	92	~~98~~
3	9	~~15~~	21	27	33	39	~~45~~	51	57	63	69	~~75~~	81	87	93	82
4	~~10~~	16	22	~~28~~	34	~~40~~	46	52	58	64	~~70~~	76	82	88	94	~~100~~
5	11	17	23	29	~~35~~	41	47	53	59	~~65~~	71	77	83	89	~~95~~	101
6	~~12~~	~~18~~	~~24~~	~~30~~	~~36~~	~~42~~	~~48~~	~~54~~	~~60~~	~~66~~	~~72~~	~~78~~	~~84~~	~~90~~	~~96~~	~~102~~

FIGURA 2 A peneira de Eratóstenes.

A PENEIRA DE ERATÓSTENES não é somente uma curiosidade histórica; ainda é um dos métodos conhecidos mais eficientes de se fazer listas extensas de primos. E métodos correlacionados têm levado a um progresso significativo sobre o que é provavelmente o mais famoso grande problema não resolvido acerca desses números: a conjectura de Goldbach. O

matemático amador alemão Christian Goldbach correspondia-se com muitas das mais ilustres figuras de seu tempo. Em 1742, afirmou uma série de conjecturas curiosas a respeito dos números primos em uma carta a Leonhard Euler. Os historiadores mais tarde notaram que René Descartes dissera muita coisa parecida alguns anos antes. A primeira das afirmações de Goldbach era: "Todo inteiro que possa ser escrito como a soma de dois primos, pode ser também escrito como a soma de quantos primos se queira, até todos os termos serem unidades." A segunda, acrescentada na margem da sua carta, dizia: "Todo inteiro maior que 2 pode ser escrito como a soma de três primos." Como a definição atual de "primo", há exceções óbvias para essas afirmativas. Por exemplo, 4 não é a soma de três primos, porque o menor primo é 2, de modo que a soma de três primos deve ser no mínimo 6. Mas na época de Goldbach, o número 1 era considerado primo. É possível refrasear diretamente suas conjecturas usando a convenção moderna.

Em sua resposta, Euler recorda uma conversa anterior com Goldbach, quando este assinalou que sua primeira conjectura seguia-se de outra mais simples, sua terceira conjectura. "Todo inteiro par é a soma de dois primos." Com a convenção predominante de que 1 é primo, essa afirmação implica também a segunda conjectura, porque qualquer número pode ser escrito, como $n + 1$ ou $n + 2$, onde n é par. Se n é a soma de dois primos, o número original é a soma de três primos. A opinião de Euler sobre a terceira conjectura foi inequívoca: "Eu encaro isso como um teorema absolutamente certo, embora não possa prová-lo." Isso resume muito bem seu status hoje.

A convenção moderna, na qual o 1 não é primo, divide as conjecturas de Goldbach em duas diferentes. A conjectura par de Goldbach afirma:

Todo inteiro par maior que 2 é a soma de dois primos.

E a conjectura ímpar de Goldbach afirma:

Todo inteiro ímpar maior que 5 é a soma de três primos.

A conjectura par implica a ímpar, mas não o inverso.[3] Vale a pena considerar ambas as conjecturas separadamente porque ainda não sabemos se alguma das duas é verdadeira. A ímpar parece ser ligeiramente

mais fácil do que a par, no sentido de que foi feito mais progresso.

Alguns cálculos rápidos verificam a conjectura par de Goldbach para números pequenos:

$4 = 2 + 2$

$6 = 3 + 3$

$8 = 5 + 3$

$10 = 7 + 3 = 5 + 5$

$12 = 7 + 5$

$14 = 11 + 3 = 7 + 7$

$16 = 13 + 3 = 11 + 5$

$18 = 13 + 5 = 11 + 7$

$20 = 17 + 3 = 13 + 7$

É fácil continuar à mão até, digamos, 1.000 ou algo em torno disso – caso você seja persistente. Por exemplo, 1.000 = 3 + 997 e 1.000.000 = 17 + 999.993. Em 1938, Nils Pipping verificou a conjectura par de Goldbach para todos os números pares até 100.000.

Também ficou claro que à medida que o número em questão vai ficando maior, tende a haver mais e mais maneiras de escrevê-lo como soma de primos. Isso faz sentido. Se você toma um número par grande, e vai subtraindo primos um de cada vez, qual é a probabilidade de *todos* os resultados serem números compostos? Basta aparecer um primo no meio da lista de diferenças resultante e a conjectura verifica-se para esse número. Utilizando recursos estatísticos dos primos, podemos avaliar a probabilidade de tal resultado. Os analistas Godfrey Harold Hardy e John Littlewood fizeram esse cálculo em 1923, e derivaram uma fórmula plausível, mas não rigorosa, para a quantidade de maneiras diferentes de expressar um dado número n como soma de dois primos: aproximadamente $n/[2(\ln n)^2]$. Essa quantidade aumenta à medida que n cresce e também concorda com a evidência numérica. Mas mesmo que esse cálculo pudesse se tornar rigoroso, poderia haver uma rara exceção ocasional, então não é grande ajuda.

O principal obstáculo para se obter uma prova da conjectura de Goldbach é que ela combina duas propriedades muito diferentes. Números primos são definidos em termos de multiplicação, mas as conjecturas

tratam de adição. Assim é extraordinariamente difícil relacionar a conclusão desejada com quaisquer características razoáveis dos primos. Parece não haver lugar algum que sirva de ponto de apoio para uma alavanca. Isso deve ter soado como música aos ouvidos da editora Faber & Faber em 2000, quando ofereceu um prêmio de 1milhão de dólares para uma prova da conjectura para promover o romance *Uncle Petros and Goldbach's Conjecture*, de Apostolos Doxiadis. O prazo final era apertado: uma solução deveria ser apresentada antes de abril de 2002. Ninguém conseguiu ter sucesso na reivindicação do prêmio, o que não é de surpreender, dado que o problema permanece sem solução há mais de 250 anos.

A CONJECTURA DE GOLDBACH é muitas vezes reformulada como uma questão sobre adição de conjuntos de inteiros. A conjectura par é o exemplo mais simples dessa forma particular de pensar, porque somamos apenas *dois* conjuntos de inteiros. Para fazer isso, pegue qualquer número do primeiro conjunto, some qualquer número do segundo conjunto e aí pegue o conjunto de todas as somas. Por exemplo, a soma de {1, 2, 3} e {4, 5} contém 1 + 4,2 + 4,3 + 4,1 + 5,2 + 5,3 + 5, que é {5, 6, 7, 8}. Alguns números ocorrem mais de uma vez, por exemplo, 6 = 2 + 4 = 1 + 5. Vou chamar esse tipo de repetição de "sobreposição".

A conjectura par de Goldbach pode agora ser reformulada: se somarmos o conjunto de primos com ele mesmo, o resultado contém todo número par maior que 2. Essa reformulação pode soar um pouco trivial – e é –, mas passa o problema para uma área em que existem alguns poderosos teoremas gerais. O número 2 é um pequeno aborrecimento, mas do qual podemos nos livrar facilmente. Ele é o único primo par, e se somarmos a ele qualquer outro primo, o resultado é ímpar. Logo, no que se refere à conjectura par de Goldbach, podemos esquecer o 2. No entanto, precisamos de 2 + 2 para representar 4, então precisamos também restringir a atenção para números pares que sejam, no mínimo, iguais a 6.

Como experimento simples, considere os números pares até 30 inclusive. Há nove números primos ímpares nesse intervalo: {3, 5, 7, 11, 13, 17, 19, 23, 29}. Somando esses números obtemos a Figura 3: marquei em negrito as somas que são inferiores ou iguais a 30 (um intervalo de números pares que inclui todos os primos até 29). Surgem dois padrões simples. A tabela toda é simétrica em relação à sua diagonal principal

porque $a + b = b + a$. Os números em negrito ocupam aproximadamente a metade superior esquerda da tabela, acima da linha (diagonal) grossa. No máximo, eles tendem a avançar além dessa linha no meio. Isso acontece porque, de modo geral, os números primos grandes são mais raros que os pequenos. A região adicional mais do que compensa os dois 32, nos cantos superior direito e inferior esquerdo.

	3	5	7	11	13	17	19	23	29
3	**6**	**8**	**10**	**14**	**16**	**20**	**22**	**26**	**32**
5	**8**	**10**	**12**	**16**	**18**	**22**	**24**	**28**	34
7	**10**	**12**	**14**	**18**	**20**	**24**	**26**	30	36
11	**14**	**16**	**18**	**22**	**24**	**28**	**30**	34	40
13	**16**	**18**	**20**	**24**	**26**	**30**	32	36	42
17	**20**	**22**	**24**	**28**	**30**	34	36	40	46
19	**22**	**24**	**26**	**30**	32	36	38	42	48
23	**26**	**28**	**30**	34	36	40	42	46	52
29	**32**	34	36	40	42	46	48	52	58

FIGURA 3 Somas de pares de números primos até 30. Negrito: somas que valem 30 ou menos. Linha grossa: diagonal. Região sombreada: eliminação de pares simetricamente relacionados. A região sombreada é ligeiramente maior que um quarto do quadrado.

Agora fazemos algumas estimativas aproximadas. Eu poderia ser mais preciso, mas essas são boas o bastante. A quantidade de casas na tabela é 9 × 9 = 81. Cerca de metade dos números nessas casas está no triângulo superior esquerdo. Devido à simetria eles aparecem em pares, exceto ao longo da diagonal – logo, a quantidade de casas não correlacionadas é cerca de $^{81}/_{4}$, aproximadamente vinte. A quantidade de inteiros ímpares no intervalo de 6 a 30 é treze. Então as vinte (e mais) somas em negrito devem abranger apenas treze números pares. Há mais somas potenciais de dois primos na região à direita do que há números pares. É como lançar vinte

bolas em treze cocos numa quermesse. Você tem uma chance razoável de acertar muitos deles. Mesmo assim, pode errar alguns. Podem estar faltando ainda alguns números pares.

Nesse caso não estão, mas esse tipo de argumento de contagem não pode eliminar tal possibilidade. No entanto, nos diz que deve haver, sim, um pouco de sobreposição, onde o mesmo número em negrito ocorre várias vezes no quarto relevante da tabela. Por quê? Porque vinte somas precisam se encaixar em apenas treze membros. Logo, em média, cada número em negrito aparece cerca de 1,5 vez. (O número real de somas é 27, então uma estimativa melhor mostra que cada número em negrito aparece duas vezes.) Se estiver faltando algum número par, a sobreposição deve ser ainda maior.

Podemos fazer o mesmo jogo com um limite superior maior – digamos, 1 milhão. Uma fórmula chamada teorema do número primo (Capítulo 9) fornece uma estimativa simples para a quantidade de primos até determinado valor x. A fórmula é $x/\ln x$. Aqui, a estimativa é de cerca de 72.380. (O número exato é 78.497.) A região sombreada correspondente ocupa cerca de um quarto da tabela, de modo que fornece por volta de $n2/4$ = 250 bilhões de números em negrito: somas de dois primos nesse intervalo. Isso é imensamente maior do que a quantidade de números pares no intervalo, que é metade de 1 milhão. Agora o tamanho da sobreposição deve ser gigantesco, com cada soma ocorrendo em média 500 mil vezes. Logo, a chance de qualquer número par escapar é bastante reduzida.

Com mais esforço, podemos transformar essa abordagem em uma estimativa da probabilidade de que algum número par num dado intervalo não seja a soma de dois primos, admitindo que eles estejam distribuídos aleatoriamente e com frequências dadas pelo teorema do número primo – isto é, algo em torno de $x/\ln x$ primos a menos em qualquer x dado. Foi isso que Hardy e Littlewood fizeram. Eles sabiam que sua abordagem não era rigorosa, porque os primos são definidos por um processo específico e na verdade não se distribuem ao acaso. Não obstante, é sensato esperar que os resultados sejam consistentes com esse modelo probabilístico, porque a propriedade que define os primos parece ter muito pouca conexão com aquilo que acontece quando somamos dois deles.

Vários métodos-padrão nessa área adotam ponto de vista semelhante, mas têm um cuidado extra de tornar o argumento rigoroso. Os métodos de

peneira, que se baseiam na peneira de Eratóstenes, são exemplos. Teoremas gerais acerca da densidade de números em somas de dois conjuntos – a proporção de números que ocorrem, à medida que os conjuntos ficam muito grandes – fornecem outras ferramentas úteis.

QUANDO UMA CONJECTURA MATEMÁTICA acaba se revelando correta, sua história frequentemente segue um padrão definido. Durante um espaço de tempo, várias pessoas provam que a conjectura é verdadeira, contanto que se apliquem restrições especiais. Cada resultado desses constitui uma melhora do anterior relaxando algumas das restrições, mas esse processo eventualmente acaba perdendo o fôlego. Por fim, uma ideia nova e muito mais sagaz completa a prova.

Por exemplo, uma conjectura na teoria dos números pode afirmar que todo inteiro positivo pode ser representado de alguma maneira usando, digamos, seis números especiais (primos, quadrados, cubos, seja o que for). Aqui as características essenciais são *todo* inteiro positivo e *seis* números especiais. Avanços iniciais levam a resultados muito mais ineficientes, mas estágios sucessivos no processo os aperfeiçoam lentamente.

O primeiro passo muitas vezes é uma prova na seguinte linha: todo inteiro positivo que não seja divisível por 3 ou por 11, exceto alguma quantidade finita deles, pode ser representado em termos de alguma quantidade gigantesca de números especiais – digamos 10^{666}. É típico o teorema não especificar quantas são as exceções, então o resultado não pode ser aplicado diretamente a qualquer inteiro específico. O passo seguinte é delimitar o território: ou seja, provar que todo inteiro maior que $10^{10^{42}}$ pode ser assim representado. Então a restrição da divisibilidade por 3 é eliminada, seguida de um avanço similar para o 11. Depois disso, sucessivos autores reduzem um dos números 10^{666} ou $10^{10^{42}}$, muitas vezes ambos. Uma melhora representativa poderia ser que todo inteiro maior que $5{,}8 \times 10^{17}$ pudesse ser representado usando, no máximo, 4.298 números especiais, por exemplo.

Enquanto isso, outros pesquisadores estão trabalhando de baixo para cima a partir de números pequenos, muitas vezes com auxílio do computador, provando que, digamos, todo número menor ou igual a 10^{12} pode ser representado usando, no máximo, seis números especiais. Em um ano, 10^{12} foi melhorado em cinco estágios, por diferentes pesquisadores ou grupos, para $11{,}0337 \times 10^{29}$. Esses aperfeiçoamentos não são nem fáceis

nem rotineiros, mas a maneira como são alcançados envolve métodos especiais intrincados que não fornecem pista alguma a respeito de uma abordagem mais genérica, e cada contribuição sucessiva é mais longa e mais complicada. Após alguns anos desse tipo de aperfeiçoamento progressivo, aplicando as mesmas ideias mas com computadores mais potentes e novas manobras, esse número subiu para 10^{43}. Mas agora o método chega a um impasse, e todos concordam que, por mais manobras que se façam, jamais se chegará à conjectura completa.

A essa altura, a conjectura desaparece de cena, porque ninguém mais está trabalhando nela. Às vezes, o progresso para totalmente. Às vezes, passam-se vinte anos sem novidade alguma… E aí, aparentemente do nada, Cheesberger e Frits anunciam que, ao reformular a conjectura em termos de quasipilhas metaergódicas complexas e aplicando a teoria colaboracionista bizantina, obtiveram uma prova completa. Após vários anos argumentando sobre pontos de lógica fina, e tapando alguns "buracos", a comunidade matemática aceita que a prova está correta, e imediatamente pergunta se há uma maneira melhor de obter o mesmo resultado, ou levá-lo mais adiante.

Você verá esse padrão se repetir muitas vezes nos próximos capítulos. Como esses relatos acabam ficando tediosos, não importa quão orgulhosos Buggins e Krumm estejam de seu mais recente aperfeiçoamento da conjectura Jekyll-Hyde de 1,773 para 1,771 + ε para qualquer ε positivo, eu descreverei algumas contribuições representativas e deixarei o resto de fora. Não para negar a importância do trabalho de Buggins e Krumm, eles podem até ter pavimentado o caminho para a grande sacada Cheesberger-Frits. Mas apenas estudiosos, seguindo a evolução da história, poderão aguardar o próximo microaperfeiçoamento com a respiração suspensa.

Futuramente darei menos detalhes, mas vamos ver como foi com Goldbach.

TEOREMAS QUE AVANÇAM parte do caminho no sentido de estabelecer a conjectura de Goldbach têm sido provados. O primeiro grande avanço veio em 1923, quando Hardy e Littlewood usaram suas técnicas analíticas para provar a conjectura ímpar de Goldbach para todos os números ímpares suficientemente grandes. No entanto, sua prova apoiava-se em outra grande conjectura, a hipótese generalizada de Riemann, que discutiremos no Capítulo 9. Esse problema ainda está em aberto, de modo que a

abordagem deles tinha uma lacuna significativa. Em 1930, Lev Schnirelmann preencheu esse vazio usando uma versão sofisticada do seu raciocínio, baseada nos métodos de peneira. Ele provou que uma proporção diferente de zero de todos os números pode ser representada como soma de dois primos. Combinando esse resultado com algumas generalidades sobre adição de sequências entre si, ele provou que existe algum número C tal que todo inteiro maior que 1 é a soma de, no máximo, C números primos. Esse valor tornou-se conhecido como constante de Schnirelmann. Ivan Matveyevich Vinogradov obteve resultados similares em 1937, mas seu método também não especificava o tamanho do "significativamente grande". Em 1939, K. Borozdin provou que não é maior que $3^{14.348.907}$. Em 2002 Liu Ming-Chit e Wang Tian-Ze haviam reduzido esse "limite superior" para $e^{3.100}$, que é aproximadamente $2 \times 10^{1.346}$. Esse é um valor muito menor, mas ainda grande demais para os números intermediários serem checados por computador.

Em 1969, N.I. Klimov obteve a primeira estimativa específica para a constante de Schnirelmann: é no máximo 6 bilhões. Outros matemáticos reduziram consideravelmente esse número, e em 1982 Hans Riesel e Robert Vaughan já o tinham diminuído para 19. Embora 19 seja bem melhor do que 6 bilhões, a evidência apontava para a constante de Schnirelmann como sendo meramente 3. Em 1995, Leszek Kaniecki reduziu o limite superior para 6, com cinco primos para qualquer número ímpar, mas precisava assumir a veracidade da hipótese de Riemann. Seus resultados combinados com a verificação numérica de J. Richstein da hipótese de Riemann até 4×10^{14} provaria que a constante de Schnirelmann é no máximo 4, mais uma vez admitindo a hipótese de Riemann. Em 1997, Jean-Marc Deshouillers, Gove Effinger, Herman te Riele e Dmitrii Zinoviev mostraram que a hipótese generalizada de Riemann (Capítulo 9) implica a conjectura ímpar de Goldbach. Ou seja, todo número ímpar exceto 1, 3 e 5 é a soma de três primos.

Considerando que a hipótese de Riemann atualmente não está provada, vale a pena remover essa premissa. Em 1995, o matemático francês Olivier Ramaré reduziu a estimativa superior para representar números ímpares até 7 sem usar a hipótese de Riemann. Na verdade, ele provou algo bem mais concreto: todo número par é uma soma de, no máximo, seis primos. (Para lidar com números ímpares subtraia 3: o resultado é par, portanto é a soma de seis ou menos primos. O número original é essa soma mais o primo 3, requerendo sete primos ou menos.) O resultado-chave de Ramaré

é que para qualquer número n maior que e^{67} (cerca de $1{,}25 \times 10^{29}$), pelo menos um quinto dos números entre n e $2n$ são a soma de dois primos. Usando métodos de peneira, em conjunção com um teorema de Hans-Heinrich Ostmann sobre somas de sequências, refinados por Deshouillers, isso conduz a uma prova de que todo número par maior que 10^3 é uma soma de, no máximo, seis primos.

O obstáculo restante é lidar com a lacuna entre 4×10^{14}, onde Jörg Richstein verificara o teorema por computador, e 10^{30}. Como é comum, os números são grandes demais para uma pesquisa direta por computador, então Ramaré provou uma série de teoremas especializados sobre a quantidade de primos em intervalos pequenos. Esses teoremas dependem da veracidade da hipótese de Riemann até limites específicos, que podem ser verificados por computador. Logo, a prova consiste principalmente em deduções conceituais de lápis e papel, com a ajuda do computador nesse aspecto particular. Ramaré acabava seu artigo ressaltando que em princípio uma abordagem similar poderia reduzir o número de primos de sete para cinco. No entanto, havia enormes obstáculos práticos, e ele escreveu que tal prova "não pode ser alcançada pelos computadores de hoje".

Em 2012, Terence Tao superou essas dificuldades com algumas ideias novas e muito diferentes. Postou um artigo na internet, que enquanto escrevo está em revisão para ser publicado. Seu principal teorema é: todo número ímpar é a soma de, no máximo, cinco primos. Isso reduz a constante de Schnirelmann para 6. Tao é renomado pela sua habilidade para resolver problemas difíceis em muitas áreas da matemática. Sua prova lança no problema diversas técnicas poderosas, e requer auxílio de computador. Se o número 5 no teorema de Tao pudesse ser reduzido para 3, a conjectura ímpar de Goldbach seria provada, e o limite para a constante de Schnirelmann, reduzido para 4. Tao desconfia que isso poderia ser possível de se fazer, embora sejam necessárias outras novas ideias.

A conjectura par de Goldbach parece ainda mais difícil. Em 1998, Deshouillers, Saouter e Te Riele a verificaram para todos os números pares até 10^{14}. Em 2007, Tomás Oliveira e Silva melhorou esse número para 10^{18}, e sua computação prossegue. Sabemos que todo inteiro par é a soma de, no máximo, seis primos – provado por Ramaré em 1995. Em 1973, Chen Jing-Run provou que todo inteiro par suficientemente grande é a soma de um primo e um semiprimo (ou primo ou produto de dois primos). Chegou perto, mas não em cima. Tao havia afirmado que a conjectura par de

Goldbach está além do alcance dos seus métodos. Somar três primos cria muito mais sobreposição nos números resultantes – no sentido discutido em relação à Figura 3 – que os dois primos necessários para a conjectura par, e os métodos de Tao e Ramaré exploram repetidamente essa característica.

Nos 2300 anos desde que Euclides provou diversos teoremas básicos a respeito dos primos, aprendemos um bocado sobre esses números fugidios, mas de importância vital. Mas aquilo que sabemos agora põe numa perspectiva árida a longa lista do que não sabemos.

Temos conhecimento, por exemplo, de que existem infinitos primos da forma $4k + 1$ e $4k + 3$; mais genericamente, que qualquer sequência aritmética[4] $ak + b$ para a e b fixos contém infinitos primos contanto que a e b não tenham fatores comuns. Por exemplo, suponhamos que $a = 18$. Então $b = 1, 5, 7, 11, 13$ ou 17. Portanto, existem infinitos primos de cada uma das formas $18k + 1$, $18k + 5$, $18k + 7$, $18k + 11$, $18k + 13$ ou $18k + 17$. Isso não se aplica a, digamos, $18k + 6$, porque 18 é múltiplo de 6. Nenhuma sequência aritmética pode conter *apenas* números primos, mas uma importante descoberta recente, o teorema de Green-Tao, mostra que o conjunto de primos contém sequências aritméticas tão longas quanto se queira. A prova, obtida em 2004 por Ben Green e Tao, é profunda e difícil. Isso nos dá esperança: questões difíceis em aberto, por mais impenetráveis que possam parecer, às vezes podem ser respondidas.

Pondo na cabeça nosso chapéu de algebrista, imediatamente nos perguntamos mais a respeito de complicadas fórmulas envolvendo k. Não existem números primos da forma k^2, e nenhum exceto 3 para $k^2 - 1$, porque essas expressões são fatoráveis. Contudo, a expressão $k^2 + 1$ não tem fatores óbvios, e aqui achamos uma profusão de primos:

$$2 = 1^2 + 1;\ 5 = 2^2 + 1;\ 17 = 4^2 + 1;\ 37 = 6^2 + 1$$

e assim por diante. Um exemplo maior sem nenhum significado específico é:

$$18.672.907.718.657 = (4.321.216)^2 + 1$$

Conjectura-se que existam infinitos primos desse tipo, mas nenhuma afirmação como essa foi ainda provada para qualquer polinômio específico no qual k ocorre numa potência maior que a primeira. Uma conjectura

muito plausível é a feita por V. Bouniakowsky em 1857: qualquer polinômio em k que não tenha divisores óbvios representa infinitos primos. As exceções aqui incluem não somente polinômios redutíveis, tais como $k^2 + k + 2$, que é sempre divisível por 2, apesar de não ter fatores algébricos.

Alguns polinômios parecem ter propriedades especiais. O caso clássico é $k^2 + k + 41$, que é primo para $k = 0, 1, 2, \ldots, 40$, e de fato também para $k = -1, -2, \ldots, -40$. Longas sequências de números primos para valores consecutivos de k são raras, e sabe-se bastante acerca delas. Mas a área toda ainda é muito misteriosa.

Quase tão famosa quanto a conjectura de Goldbach, e aparentemente tão difícil, é a conjectura dos números primos gêmeos consecutivos: existem infinitos pares de primos cuja diferença é 2. Alguns exemplos são:

3, 5 5, 7 11, 13 17, 19

O maior par conhecido de números primos gêmeos (até janeiro de 2012) era:

$$3.756.801.695.685 \times 2^{666.669} \pm 1$$

que tem 200.700 dígitos decimais. Foram encontrados pelo projeto de computação distribuída PrimeGrid em 2011. Em 1915, Viggo Brun usou uma variante da peneira de Eratóstenes para provar que a soma dos inversos de todos os números primos gêmeos converge, diferentemente da soma dos inversos de todos os primos. Assim, nesse sentido, números primos gêmeos são relativamente raros. Ele provou também, usando métodos semelhantes, que existem infinitos inteiros n tais que n e $n + 2$ têm, no máximo, nove fatores primos. Hardy e Littlewood usaram seus métodos heurísticos para argumentar que a quantidade de pares de números primos gêmeos menores que x deveria ser assintótica a

$$2a \frac{n}{(\ln n)^2}$$

onde a é uma constante cujo valor é aproximadamente 0,660161. A ideia subjacente é que, para esse propósito, pode-se presumir que os primos surjam aleatoriamente, numa taxa que torna a quantidade de primos até x aproximadamente igual a $x/\ln x$. Há muitas conjecturas e fórmulas heurísticas semelhantes, porém, mais uma vez, não há provas rigorosas.

De fato, há centenas de questões em aberto em relação aos números

primos. Umas são apenas curiosidades; outras, profundas e significativas. Destas últimas, encontraremos algumas delas no Capítulo 9. Apesar de todos os avanços feitos pelos matemáticos nos últimos dois milênios e meio, os humildes primos não perderam nada do seu fascínio nem do seu mistério.

3. O quebra-cabeça de pi

A quadratura do círculo

NÚMEROS PRIMOS SÃO uma ideia antiga, mas os círculos são mais antigos ainda. Eles conduziram a um grande problema que levou mais de 2 mil anos para ser resolvido. É um entre vários problemas geométricos correlatos que chegaram a nós da Antiguidade. O personagem principal da história é o número π (a letra grega "pi"), que conhecemos na escola em conexão com círculos e esferas. Numericamente ele vale 3,14159 e um pouquinho mais; com frequência é usada a aproximação $^{22}/_{7}$. Os dígitos de π não param nunca, e nunca repetem a mesma sequência. O atual recorde para o cálculo dos dígitos de π é 10 trilhões de dígitos, de Alexander Yee e Chigeru Kondo em outubro de 2011.[1] Cálculos como esses são significativos como meio de testar computadores rápidos, ou para inspirar e testar novos métodos de calcular π, mas pouca coisa advém dos resultados numéricos. A razão do interesse em π não é para calcular o comprimento de uma circunferência em torno de um círculo. O mesmo número estranho aparece em toda a matemática, não só em fórmulas relacionadas a círculos e esferas, e nos leva a águas verdadeiramente profundas. De qualquer modo, as fórmulas aprendidas na escola são importantes e refletem as origens de π na geometria grega.

Ali, um dos grandes problemas era a tarefa não solucionada de quadrar um círculo. Essa expressão é frequentemente empregada de maneira coloquial para indicar uma abordagem errada para alguma coisa, mais ou menos como tentar encaixar um pino quadrado num furo redondo. Como em muitas outras expressões comuns extraídas da ciência, o significado dessa mudou com os séculos.[2] Nos tempos da Grécia antiga, tentar quadrar um círculo era uma ideia perfeitamente razoável. A diferença nas duas formas – reta ou curva – é totalmente irrelevante: problemas similares têm soluções válidas.[3] No entanto, ocorreu que este problema em particular não pode ser resolvido usando os métodos especificados. A prova é engenhosa e técnica, mas sua natureza geral é compreensível.

Em matemática, quadrar o círculo significa construir um quadrado cuja *área* seja igual à de um círculo dado, usando os métodos tradicionais de

Euclides. A geometria grega na verdade permitia outros métodos, então um dos aspectos do problema é especificar que métodos devem ser usados. A impossibilidade de solucionar o problema é, portanto, uma constatação das limitações desses métodos; não implica que não possamos determinar a área de um círculo. Simplesmente precisamos encontrar outra abordagem. A prova da impossibilidade explica por que os geômetras gregos e seus sucessores fracassaram em encontrar uma construção como a exigida: tal construção não existe. Em retrospecto, isso explica por que tiveram de introduzir métodos mais esotéricos. Assim a solução, apesar de ser negativa, esclarece o que de outra maneira seria um quebra-cabeça histórico. E também impede as pessoas de perder tempo numa contínua busca por uma construção que não existe – com exceção de algumas almas penadas que lamentavelmente parecem incapazes de receber a mensagem, não importa com quanto cuidado seja explicada.[4]

Nos *Elementos* de Euclides, os métodos tradicionais para construir figuras geométricas são versões idealizadas de dois instrumentos matemáticos: a régua e o compasso. Para ser pedante, compasso*s*, mas seguirei o jeito comum de falar, evitando o plural. Esses instrumentos são usados para "desenhar" diagramas numa folha de papel para anotações, o plano euclidiano.

A forma dos instrumentos determina o que podem desenhar. Um compasso é constituído de duas hastes rígidas, presas por uma articulação. Uma delas tem uma ponta afilada, a outra segura um lápis bem apontado. O instrumento é usado para desenhar um círculo, ou parte de um círculo, com um centro específico e um raio específico. Uma régua é mais simples: possui uma borda reta e é usada para desenhar uma linha reta. Ao contrário das réguas que você compra nas papelarias, a de Euclides não tem marcas, e essa é uma restrição importante para a análise matemática do que ela é capaz de criar.

O sentido em que a régua e o compasso do geômetra são idealizações é claro e direto: presume-se que se desenhem linhas infinitamente finas. Mais ainda, as linhas retas são exatamente retas e os círculos são perfeitamente redondos. O papel é perfeitamente plano e regular. O outro elemento-chave na geometria de Euclides é a noção de ponto, outra idealização. Um ponto é uma pequena marca no papel, mas é uma impossibilidade física: ele não tem tamanho. "Um ponto", diz Euclides na

primeira sentença dos *Elementos*, "é aquilo que não tem parte." Isso nos faz lembrar um átomo, ou se você tem alguma noção de física moderna, uma partícula subatômica, mas em comparação com um ponto geométrico, estas são gigantes. Da perspectiva humana cotidiana, porém, o ponto ideal de Euclides, um átomo, e uma marquinha de lápis numa folha de papel são suficientemente similares para os propósitos da geometria.

Essas idealizações não são possíveis de ser obtidas no mundo real, por mais cuidado que se tenha ao fabricar os instrumentos e apontar o lápis, e por mais liso que seja o papel. Mas a idealização pode ser uma virtude, porque essas exigências tornam a matemática muito mais simples. Por exemplo, duas linhas de lápis cruzam-se numa pequena região imprecisa na forma de um paralelogramo, mas as linhas matemáticas se encontram num único ponto. Percepções adquiridas de círculos e retas muitas vezes podem ser transferidas para retas e círculos reais, imperfeitos. É assim que a matemática faz a sua mágica.

Dois pontos determinam uma reta, a única que passa por eles. Para construir a reta, coloque a sua régua ideal de maneira que ela passe pelos dois pontos e corra seu lápis ideal ao longo dela. Dois pontos também determinam um círculo: escolha um deles como centro e coloque ali a ponta-seca do compasso; então ajuste o instrumento de modo que a ponta do lápis caia em cima do outro ponto. Agora gire o lápis em volta da outra haste num arco, mantendo fixo o ponto central. Duas retas determinam um único ponto no lugar onde se cruzam, a menos que sejam paralelas, e nesse caso não se cruzam, mas aqui existe à espera uma caixa de Pandora repleta de questões lógicas. Uma reta e uma circunferência – o contorno do círculo – determinam dois pontos, caso se cruzem; um ponto, se a reta for tangente ao círculo, e nenhum se o círculo for pequeno demais para se encontrar com a reta. Da mesma maneira, dois círculos cruzam-se em dois pontos, em um ou nenhum.

Distância é um conceito fundamental no moderno tratamento da geometria euclidiana. A distância entre dois pontos quaisquer é medida ao longo da reta que os liga. Euclides conseguiu fazer sua geometria funcionar sem um conceito explícito de distância, encontrando um jeito de dizer que dois segmentos de reta têm o *mesmo* comprimento sem definir comprimento em si. Na verdade, isso é fácil: basta abrir o compasso entre as extremidades de um segmento, transferi-lo para o segundo e ver se as extremidades se encaixam. Se isso ocorre, os comprimentos são iguais; caso contrário, não são. Em nenhuma etapa chegamos a medir o comprimento

real.

A partir desses ingredientes básicos, os geômetras podem compor formas e configurações mais interessantes. Três pontos determinam um triângulo, a menos que estejam todos sobre a mesma reta. Quando duas retas se cruzam, formam um ângulo. Um ângulo reto é especialmente significativo; uma linha reta corresponde a dois ângulos retos dispostos sucessivamente. E assim por diante. Os *Elementos* de Euclides consiste em treze livros, que investigam consequências cada vez mais profundas desse início simples.

O grosso dos *Elementos* é constituído de teoremas – traços válidos da geometria. Mas Euclides também explica como solucionar problemas geométricos, usando "construções" baseadas em régua e compasso. Dados dois pontos unidos por um segmento de reta, achar seu ponto médio. Ou trisseccionar o segmento: achar um ponto exatamente a um terço do percurso total. Dado um ângulo, construir outro que o bissecsione – tenha metade do seu tamanho. Mas algumas construções simples revelaram-se ardilosas. Dado um ângulo, construir outro que o trisseccione – tenha um terço do seu tamanho. É possível fazer isso com um segmento de reta, mas ninguém conseguiu achar um método para ângulos. Aproximações, tão próximas quanto se deseje, sim. Construções exatas usando uma régua sem marcação e um compasso, não. Todavia, de qualquer maneira ninguém precisa trisseccionar exatamente um ângulo, de modo que essa questão específica não causou muito problema.

Mais constrangedora era uma construção que não podia ser ignorada: dado um círculo, construir um quadrado que tenha a mesma área. Este é o problema da quadratura do círculo. Do ponto de vista grego, se você não pudesse resolvê-lo, não tinha o direito de alegar que o círculo *tivesse* área. Mesmo que ele visivelmente englobe um espaço bem-definido, e que intuitivamente a área é *quanto* espaço está englobado. Euclides e seus sucessores, sobretudo Arquimedes, contentaram-se com uma solução pragmática: assumir que os círculos têm área, mas não esperar que alguém seja capaz de construir quadrados com a mesma área. Ainda assim, pode-se dizer muita coisa; por exemplo, pode-se provar, com pleno rigor lógico, que a área de um círculo é proporcional ao quadrado do seu diâmetro. O que não se pode fazer, sem quadrar o círculo, é construir um segmento de reta cujo comprimento seja essa constante de proporcionalidade.

Os gregos não conseguiram quadrar o círculo usando régua e compasso, então recorreram a outros métodos. Um deles usou uma curva chamada

quadratriz.[5] A importância atribuída por eles ao uso apenas da régua e do compasso foi exagerada por alguns comentaristas posteriores, e nem mesmo está claro que os gregos consideravam a quadratura do círculo um assunto vital. Por volta do século XIX, porém, o problema estava se tornando um grande aborrecimento. Matemáticos incapazes de responder a uma questão tão direta eram como um chef de cozinha conceituado que não sabe preparar um ovo cozido.

A QUADRATURA DO CÍRCULO soa como um problema na geometria. Isso ocorre porque é um problema de geometria. Mas sua solução acabou residindo não na geometria, mas na álgebra. Fazer ligações inesperadas entre áreas da matemática aparentemente não relacionadas muitas vezes está no cerne da resolução de um grande problema. Aqui, a ligação não foi totalmente inesperada, sem precedentes, mas o elo com a quadratura do círculo não foi apreciado de início. E mesmo quando foi, havia uma dificuldade técnica, e lidar com ela exigia ainda outra área da matemática: a análise, a versão rigorosa do cálculo. Ironicamente, a primeira grande descoberta veio de uma quarta área: a teoria dos números. E solucionou um problema geométrico para o qual os gregos, nem nos seus sonhos mais doidos, acreditariam possuir uma solução, e, até onde podemos dizer, jamais pensaram: como construir, com régua e compasso, um polígono regular de dezessete lados.

Parece loucura, especialmente se acrescentarmos que tal construção não existe para polígonos regulares com 7, 9, 11, 13 ou 14 lados, mas a fazemos para 3, 4, 5, 6, 8, 10 e 12 lados. Todavia, há método por trás da loucura, e é o método que enriqueceu a matemática.

Primeiro: o que é um polígono regular? Um polígono é uma forma delimitada por segmentos de reta. Ele será regular se esses segmentos tiverem o mesmo comprimento e se encontrarem em ângulos iguais. O exemplo mais familiar é o quadrado: todos os quatro lados têm o mesmo comprimento e os quatro ângulos são retos. Existem outras formas com quatro lados iguais ou com quatro ângulos iguais: o losango e o retângulo, respectivamente. Apenas o quadrado tem ambas as características. Um polígono regular de três lados é um triângulo equilátero, um polígono regular de cinco lados é um pentágono regular, e assim por diante (Figura 4). Euclides fornece construções régua e compasso para polígonos regulares de 3, 4 e 5 lados. Os gregos também sabiam como duplicar

repetidamente o número de lados, dando 6, 8, 10, 12, 16, 20, e assim por diante. Combinando as construções de polígonos regulares de três e cinco lados, podiam obter um polígono de quinze lados. Mas aí terminava o conhecimento deles. E durante cerca de 2 mil anos foi assim que as coisas ficaram. Ninguém imaginava que algum outro número fosse viável. Ninguém sequer perguntava, simplesmente parecia óbvio que nada mais podia ser feito.

FIGURA 4 Os primeiros polígonos regulares. *Da esquerda para a direita*: triângulo equilátero, quadrado, pentágono, hexágono, heptágono, octógono.

Foi preciso um dos maiores matemáticos que já viveu para pensar o impensável, perguntar o imperguntável e descobrir uma resposta verdadeiramente assombrosa. Ou seja, Carl Friedrich Gauss. Ele nasceu em uma família pobre, da classe operária, na cidade de Braunschweig (Brunswick), na Alemanha. Sua mãe, Dorothea, não sabia ler nem escrever, e não conseguiu anotar sua data de nascimento, mas lembrava-se de que tinha sido em uma quarta-feira, oito dias antes da festa da ascensão, em 1777. Mais tarde, Gauss calculou a data exata a partir de uma fórmula matemática que criou para a data da Páscoa. Seu pai, Gebhard, vinha de uma família de agricultores, mas ganhava a vida numa série de serviços de nível inferior: jardineiro, trabalhador na construção de canais, açougueiro de rua, servente de agência funerária. O filho era um menino-prodígio com reputação de ter corrigido, aos três anos de idade, a aritmética do pai. Suas habilidades, que se estendiam a línguas, bem como à matemática, levaram o duque de Braunschweig a financiar seus estudos no Collegium Carolinum. Como aluno de graduação, Gauss redescobriu de forma independente diversos teoremas matemáticos importantes que haviam sido provados por gente ilustre como Euler. Mas seu teorema referente ao polígono de dezessete lados chegou como um raio vindo do céu.

A essa altura, o estreito elo entre geometria e álgebra já fora entendido havia 140 anos. Em um apêndice ao *Discurso do método*, René Descartes formalizou uma ideia que vinha flutuando por algum tempo em forma rudimentar: a noção de um sistema de coordenadas. Com efeito, esse

sistema pega o plano árido de Euclides, uma folha de papel em branco, e o transforma num papel quadriculado, que engenheiros e cientistas chamam de papel para gráficos. Desenhe nesse papel duas linhas retas, uma horizontal e outra vertical: essas duas retas são chamadas eixos. Agora você pode estabelecer a localização de qualquer ponto do plano perguntando a que distância ele se encontra ao longo do eixo horizontal e a que altura do eixo vertical (Figura 5, esquerda). Esses dois números, que podem ser positivos ou negativos, fornecem uma descrição completa do ponto, e são chamados coordenadas.

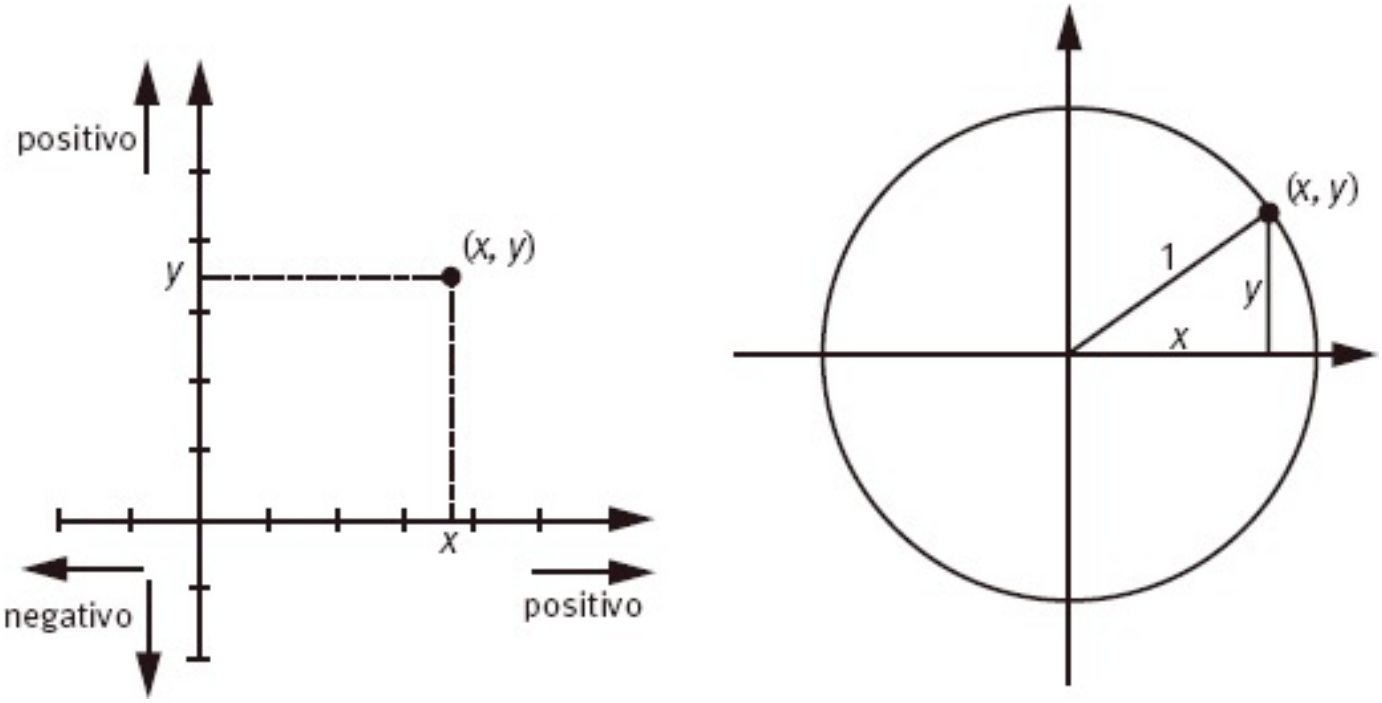


FIGURA 5 *Esquerda*: Coordenadas no plano. *Direita*: Como deduzir a equação para o círculo unitário.

Todas as propriedades geométricas de pontos, retas, círculos, e assim por diante, podem ser traduzidas em expressões algébricas com as correspondentes coordenadas. É muito difícil falar significativamente dessas correlações sem usar alguma álgebra efetiva – do mesmo modo que é difícil falar com sensatez de futebol sem mencionar a palavra “gol”. Assim, as próximas páginas vão incluir algumas fórmulas. Elas estão aí para assegurar que os principais atores envolvidos no drama tenham nomes e que a relação entre eles esteja clara. “Romeu” é muito mais fácil de acompanhar do que “o filho de um patriarca italiano que se apaixona pela linda filha do inimigo jurado de seu pai”. Nosso Romeu receberá o prosaico nome de *x*, e sua Julieta será *y*.

A título de exemplo de como a geometria se converte em álgebra, a

Figura 5 (direita) mostra como achar a equação de um círculo de raio unitário com centro na origem, onde os dois eixos se cruzam.[b] O ponto marcado tem coordenadas (*x, y*), de modo que o triângulo retângulo na figura tem lado horizontal de comprimento *x* e lado vertical de comprimento *y*. O lado maior do triângulo é o raio do círculo, que é 1. O teorema de Pitágoras nos diz que a soma dos quadrados das duas coordenadas é 1. Em símbolos, um ponto de coordenadas *x* e y encontra-se sobre o círculo se (e somente se) satisfaz a condição $x^2 + y^2 = 1$. Esta caracterização simbólica do círculo é breve e precisa, e mostra que estamos realmente falando de álgebra. Inversamente, qualquer propriedade algébrica de pares de números, qualquer equação envolvendo *x* e *y* pode ser reinterpretada como afirmações geométricas sobre pontos, retas, círculos e curvas mais elaboradas.[6]

AS EQUAÇÕES BÁSICAS da álgebra envolvem polinômios, combinações de potências de uma grandeza desconhecida *x*, a incógnita, onde cada potência é multiplicada por algum número, chamado coeficiente. A maior potência que ocorre é o grau do polinômio. Por exemplo, a equação

$$x^4 - 3x^3 - 3x^2 + 15x - 10 = 0$$

envolve um polinômio que começa com x^4, logo, seu grau é 4. Os coeficientes são 1, −3, −3, 15 e –10. Existem quatro soluções distintas: $x = 1$, 2, $\sqrt{5}$ e $\sqrt{5}$. Para esses números o termo esquerdo da equação é igual a zero – o termo direito. Polinômios de grau 1, como $7x + 2$, são ditos lineares e envolvem apenas a primeira potência da incógnita. Equações de grau 2, como $x^2 - 3x + 2$, são ditas quadráticas, e envolvem a segunda potência – o quadrado. A equação de um círculo envolve uma segunda variável, *y*. No entanto, se conhecermos uma segunda equação relacionando *x* e *y*, por exemplo a equação que define uma linha reta, então podemos resolver para *y* em termos de *x* e reduzir a equação do círculo para outra que envolva somente *x*. Essa nova equação nos diz onde a reta encontra o círculo. Nesse caso a nova equação é quadrática, com duas soluções; é assim que a álgebra reflete a geometria, onde uma reta cruza o círculo em dois pontos distintos.

Essa característica da álgebra tem uma implicação importante para construções régua e compasso. Uma construção dessas, por mais complicada que seja, divide-se em uma sequência de passos simples. Cada passo produz novos pontos em locais onde duas retas, dois círculos ou uma

reta e um círculo se cruzam. Essas retas e esses círculos são determinados por pontos anteriormente construídos. Traduzindo geometria em álgebra, pode-se provar que a equação algébrica correspondente à intersecção de duas retas é sempre linear, enquanto para uma reta e um círculo, ou dois círculos, é quadrática. Em última análise, isso acontece porque a equação de um círculo envolve x^2, mas nenhuma potência superior de x. Logo, cada passo individual numa construção corresponde, apenas, a resolver uma equação de grau 1 ou 2 .

Construções mais complexas são sequências dessas operações básicas, e um certo volume de técnica algébrica permite-nos deduzir que cada coordenada de qualquer ponto que pode ser obtido com régua e compasso é uma solução de uma equação polinomial, com coeficientes inteiros, cujo grau é uma potência de 2. Isto é, o grau deve ser um desses números: 1, 2, 4, 8, 16, e assim por diante.[7] Essa condição é necessária para existir uma construção, mas pode ser reforçada numa caracterização precisa de quais polígonos regulares são possíveis de ser construídos. De repente, uma condição algébrica ordenada emerge de uma complicada desordem geométrica – e se aplica a *qualquer construção*. Você não precisa nem saber qual construção é: apenas que ela utiliza somente régua e compasso.

Gauss estava ciente dessa elegante ideia. E sabia também (de fato, qualquer matemático competente rapidamente perceberia) que a questão de quais polígonos regulares podem ser construídos com régua e compasso se reduz a um caso especial, quando o polígono tem um número primo de lados. Para ver o porquê, pense em um número composto como 15, que é 3 × 5. Qualquer construção hipotética de um polígono regular de quinze lados automaticamente inclui a de um polígono de três lados (considere cada quinto vértice) e outra de um polígono de cinco lados (considere cada terceiro vértice) (Figura 6). Com um pouco mais de esforço você pode combinar as construções de um triângulo e de um pentágono para obter um 15-ágono, um pentadecágono.[8] Os números 3 e 5 são primos, e a mesma ideia aplica-se genericamente. Assim, Gauss concentrou-se em polígonos com um número primo de lados, e indagou qual seria a aparência da equação relevante. A resposta foi surpreendentemente clara. Construir um polígono regular de cinco lados, por exemplo, equivale a resolver a equação $x^5 - 1 = 0$. Substitua 5 por qualquer outro número primo e a expressão correspondente é verdadeira.

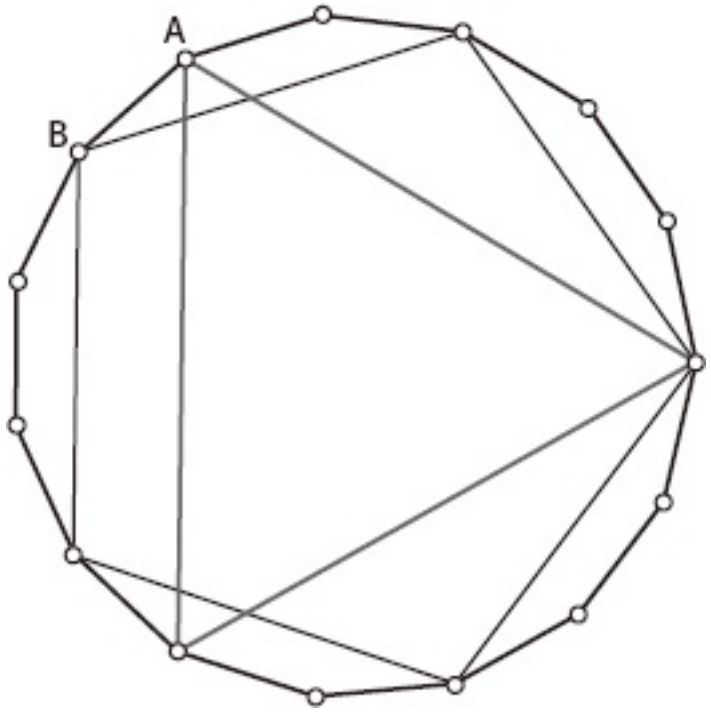


FIGURA 6 Construção de um triângulo equilátero e de um pentágono regular a partir de um 15-ágono. Para o processo inverso, observe que A e B são pontos consecutivos no 15-ágono regular.

O grau desse polinômio é 5, que *não* é uma das potências de 2 que eu listei; mesmo assim existe uma construção. Gauss logo descobriu por quê: a equação divide-se em duas partes, uma de grau 1 e outra de grau 4. Tanto 1 como 4 são potências de 2, e descobrimos que a equação de grau 4 é a crucial. Para ver o porquê, precisamos relacionar a equação com a geometria. Isso envolve um novo tipo de número, que costuma ser ignorado na matemática escolar mas é indispensável para qualquer coisa além daí. São os chamados números complexos, e a característica que os define é que no sistema dos números complexos o número −1 tem raiz quadrada.[9]

Um número "real" comum é positivo ou negativo, mas em ambos os casos seu quadrado é positivo, então −1 não pode ser quadrado de nenhum número real. Esse é um transtorno tão grande que os matemáticos inventaram um novo tipo de número "imaginário" cujo quadrado é −1. Eles precisavam de um novo símbolo para esse número, então o chamaram de i (de "imaginário"). As operações habituais da álgebra – adição, subtração, multiplicação e divisão – levam a novas combinações de números reais e imaginários tais como 3 + 2i. Esses números são ditos complexos, o que não significa "complicados", mas indica que eles vêm em duas partes: 3 e 2i. Os números reais jazem na famosa reta numerada, como os números numa régua. Os números complexos encontram-se no plano numérico, no qual a

régua imaginária é colocada em ângulo reto com a régua real, e as duas juntas formam um sistema de coordenadas (Figura 7, esquerda).

Nos últimos duzentos anos, os matemáticos têm considerado os números complexos fundamentais para o assunto. Agora reconhecemos que logicamente eles estão no mesmo patamar que os números "reais" mais familiares – que, como todas as estruturas matemáticas, são conceitos abstratos, não coisas físicas reais. Os números complexos estavam em uso difundido antes da época de Gauss, mas seu status ainda era misterioso até que ele e vários outros os desmistificaram. A fonte de sua atração era paradoxal: apesar do mistério cercando seu significado, os números complexos eram muito mais bem-comportados do que os reais. Eles supriam um ingrediente que faltava, do qual os números reais careciam. E forneciam um conjunto completo de soluções para uma equação algébrica.

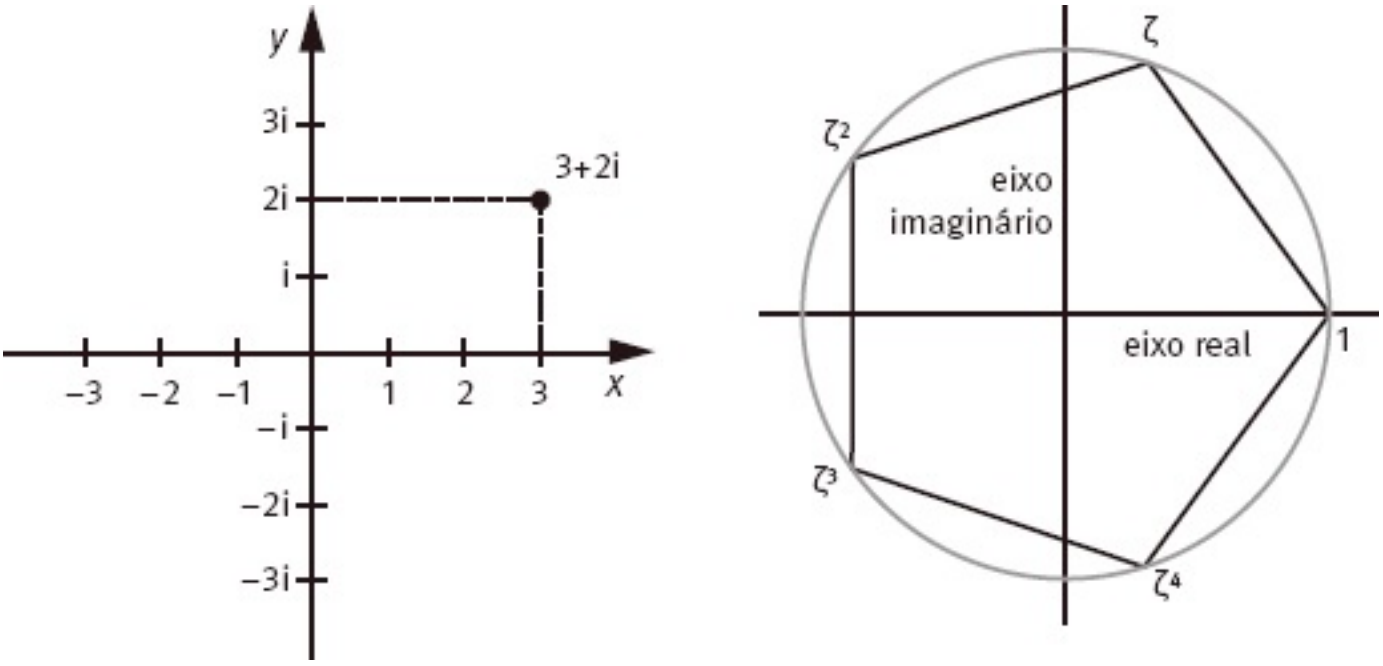


FIGURA 7 *Esquerda*: O plano complexo. *Direita*: As raízes quintas complexas da unidade.

Equações quadráticas são o exemplo mais simples. Algumas delas têm duas soluções reais, enquanto outras não têm uma sequer. Por exemplo, $x^2 - 1 = 0$ tem as soluções 1 e −1, mas $x^2 + 1 = 0$ não tem soluções. Entre as duas está $x^2 = 0$, cuja única solução é 0, mas existe um sentido no qual esta é a mesma solução "repetida duas vezes".[10] No entanto, se permitirmos soluções complexas, então $x^2 + 1 = 0$ também tem duas soluções: i e −i. Gauss não tinha escrúpulos em usar números complexos; na verdade, sua tese de doutorado forneceu a primeira prova que soava logicamente sólida do teorema fundamental da álgebra: o número de soluções complexas de

qualquer equação polinomial (com as multiplicidades corretamente contadas) é igual ao grau da equação. Assim, as quadráticas (grau 2) sempre têm duas soluções complexas, as cúbicas (grau 3) têm três soluções complexas, e assim por diante.

A equação $x^5 - 1 = 0$, que eu alego definir um pentágono regular, tem grau 5. Portanto, tem cinco soluções complexas. Existe apenas uma solução real: $x = 1$. E quanto às outras quatro? Elas fornecem quatro vértices de um pentágono regular perfeito no plano complexo, com $x = 1$ sendo o quinto (Figura 7, direita). Esta correspondência é um exemplo de beleza matemática: uma forma geométrica elegante torna-se uma equação elegante.

Agora, a equação cujas soluções são esses cinco pontos tem grau 5, que não é potência de 2. Mas, como mencionado antes, a equação de grau 5 divide-se em duas partes de graus 1 e 4, chamadas de fatores irredutíveis:

$$x^5 - 1 = (x - 1)\,(x^4 + x^3 + x^2 + x + 1)$$

("Irredutíveis" significa que não existem mais fatores, exatamente como os números primos.) O primeiro fator contém a solução real $x = 1$. O outro fator contém as quatro soluções complexas e os outros quatro vértices do pentágono. Assim tudo faz muito mais sentido, e é muito mais elegante, quando usamos números complexos.

FREQUENTEMENTE É DIFÍCIL reconstituir como os matemáticos do passado chegaram a novas descobertas, porque eles tinham o hábito de apresentar apenas o resultado final de suas deliberações, não os muitos passos em falso que tinham dado ao longo do caminho. Esse problema é muitas vezes agravado, porque os padrões naturais de pensamento em épocas passadas eram diferentes dos de hoje. Gauss, em particular, era notório por encobrir seus rastros e publicar apenas sua análise final, altamente lapidada. Mas quando se trata da pesquisa de Gauss sobre o polígono de dezessete lados, estamos em solo bastante seguro; a análise que ele acabou publicando fornece várias pistas proveitosas.

Seu ponto de partida não era novo. Diversos matemáticos anteriores estavam bem cientes de que a análise dos pentágonos regulares acima funciona de modo totalmente genérico. Construir um polígono com qualquer número n de lados equivale a resolver a equação $x^n - 1 = 0$ em números complexos. Além disso, esse polinômio é fatorável em:

$$(x - 1)\,(x^{n-1} + x^{n-2} + \ldots + x^2 + x + 1)$$

Mais uma vez o primeiro fator dá a solução real $x = 1$ e as restantes $n - 1$ soluções provêm do segundo fator. Quando n é ímpar, elas são todas complexas; quando n é par, uma delas é uma segunda solução real $x = -1$.

O que Gauss notou, e todos os outros não perceberam, é que às vezes o segundo fator pode ser expresso usando uma série de equações quadráticas. Não sendo representado como produto de fatores mais simples, pois isso não é possível, mas usando equações cujos coeficientes solucionem outras equações. O fator-chave aqui – o ponto fraco do problema – é uma propriedade elegante das equações algébricas, que surge quando resolvemos várias delas dessa maneira, uma de cada vez. O cálculo é sempre equivalente a resolver uma única equação, mas o grau geralmente aumenta. Assim, o preço que pagamos para ter menos equações é um aumento no grau. Pode ser complicado, mas há uma característica que podemos prever: até onde o grau cresce. Basta multiplicar os graus dos sucessivos polinômios.

Se forem todos quadráticos, o resultado será $2 \times 2 \times \ldots \times 2$, uma potência de 2. Então $n - 1$ deve ser uma potência de 2, se a construção existir. No entanto, essa condição nem sempre é suficiente. Quando $n = 9$, $n - 1 = 8$, que é uma potência de 2. Mas Gauss descobriu que não existe construção para um 9-ágono (eneágono) regular. A razão é que 9 não é primo.[11] Que tal o caso seguinte, no qual resolvemos uma série de quatro equações quadráticas? Agora o grau $n - 1$ da equação única correspondente é $2 \times 2 \times 2 \times 2 = 16$. Então $n = 17$, e este é primo.

A essa altura Gauss deve ter percebido que estava no caminho certo, mas há um ponto técnico adicional, possivelmente fatal. Ele havia se convencido que para que exista a construção de um polígono regular com um número primo n de lados, esse primo deve ser uma potência de 2, mais 1. Logo, essa condição é necessária para que a construção exista: se falhar, não existe tal construção. No entanto, a condição pode não ser suficiente: na verdade, existe uma profusão de equações de grau 16 que não se reduzem a uma série de quatro quadráticas.

Contudo, havia uma razão para ser otimista: as construções gregas. Que números primos ocorriam ali? Apenas três: 2, 3 e 5. São todos uma potência de 2, mais 1, ou seja, $2^0 + 1$, $2^1 + 1$ e $2^2 + 1$. A álgebra associada ao pentágono fornece outras pistas. Repassando tudo isso, Gauss provou que o polinômio de grau 16 associado ao polígono de dezessete lados pode ser

efetivamente reduzido a uma série de quadráticas. Portanto, deve existir uma construção régua e compasso. Um método similar provou que o mesmo é verdade sempre que o número de lados é um primo uma unidade maior que uma potência de 2. As ideias são um tributo à habilidade de Gauss de compreender os padrões matemáticos. No cerne dessas ideias há alguns teoremas gerais da teoria dos números, na qual não vou falar aqui. A questão é: nada disso foi acidental. Havia motivos estruturais sólidos para que funcionasse. Bastava ser um Gauss para notá-los.

Gauss não forneceu uma construção explícita, mas deu, sim, uma fórmula para as soluções da equação de grau 16 que pode ser transformada em tal construção caso você realmente queira.[12] Quando anotou suas ideias no *Disquisitiones Arithmeticae*, ele omitiu um bocado de detalhes, mas afirmou que possuía realmente provas completas. Sua descoberta épica o convenceu de que devia dedicar sua vida à matemática e não às línguas. O duque continuou a apoiar Gauss financeiramente, mas este desejava algo mais permanente e seguro. Quando o astrônomo Giuseppe Piazzi descobriu o primeiro asteroide, Ceres, só foi possível fazer umas poucas observações antes que esse novo mundo se tornasse invisível contra o brilho do Sol. Os astrônomos ficaram preocupados com a possibilidade de não serem capazes de encontrá-lo novamente. Num *tour de force* que envolveu novas técnicas para calcular órbitas, Gauss predisse onde ele reapareceria – e acertou. Isso o levou a ser indicado a professor de astronomia e diretor do Observatório de Göttingen. Ele manteve o posto para o resto da vida.

O que se descobriu foi que 17 não é o único número novo desse tipo. Mais dois são conhecidos: $2^8 + 1 = 257$ e $2^{16} + 1 = 65.537$. (Um pouco de álgebra mostra que a potência à qual 2 é elevado deve ser ela própria uma potência de 2; se não, o número não pode ser primo.) Porém, o padrão é interrompido aí, pois $2^{32} + 1 = 4.294.967.297$, que é igual a $641 \times 6.700.417$, portanto não primo. Os chamados números de Fermat $2^{2^n} + 1$ são conhecidos por não serem primos para n = 5, 6, 7, ... até 32. Muitos números de Fermat maiores também são conhecidos por não serem primos. Não foram encontrados outros números primos de Fermat, mas sua existência não é de forma alguma impossível.[13] Uma construção para o polígono de 257 lados é conhecida. Um matemático dedicou muitos anos para o polígono de 65.537 lados, uma tarefa um tanto sem sentido, e, de qualquer maneira, seus resultados contêm erros.[14]

A CONCLUSÃO DA ANÁLISE de Gauss é que um polígono regular pode ser construído com régua e compasso se, e somente se, o número de lados é produto de uma potência de 2 e número primo ímpar de Fermat *distinto.* Em particular, um polígono regular de nove lados não pode ser construído dessa maneira, isso implica imediatamente que pelo menos um ângulo não pode ser trisseccionado, porque o ângulo num triângulo equilátero é sessenta graus, e um terço dele é vinte graus. Dado esse ângulo, é fácil construir um polígono regular de nove lados. Já que isso é impossível, não há construção geral com régua e compasso para trisseccionar um ângulo.

Gauss omitiu muitos detalhes das provas quando escreveu seus resultados, e os matemáticos não podiam simplesmente aceitar sua palavra. Em 1837, o matemático francês Pierre Wantzel publicou uma prova completa da caracterização feita por Gauss dos polígonos regulares passíveis de serem construídos, e deduziu a impossibilidade de trisseccionar um ângulo genérico mediante construção com régua e compasso. E também provou que é impossível construir um cubo cujo volume seja o dobro de um cubo dado, outro problema da Grécia antiga conhecido como "duplicação do cubo".

Tanto a trissecção do ângulo como a duplicação do cubo revelam-se impossíveis porque os comprimentos envolvidos satisfazem equações *cúbicas* – grau 3 – irredutíveis. Uma vez que 3 não é potência de 2, isso leva a equação a nocaute. No entanto, esse método parecia não funcionar para o problema da quadratura do círculo por motivos interessantes. Um círculo de raio unitário tem área π e o quadrado com essa área tem lado $\sqrt{\pi}$. A construção geométrica para a raiz quadrada existe, e o mesmo ocorre com a construção do quadrado, de modo que a quadratura do círculo se reduz a começar com um segmento de reta de comprimento 1 e construir um de comprimento π. Se por acaso π satisfizesse uma equação cúbica irredutível – ou qualquer equação irredutível cujo grau não seja potência de 2 –, então os métodos de Wantzel provariam que é impossível quadrar o círculo.

No entanto, ninguém conhecia nenhuma equação algébrica que fosse satisfeita exatamente por π, muito menos uma cujo grau não seja potência de 2. O valor de $^{22}/_{7}$ usado na escola satisfaz $7x - 22 = 0$, mas esta é apenas uma aproximação de π, um pouquinho grande demais, então não adianta. Se fosse possível provar que tal equação não existe – e muitos suspeitavam disso com base em que, se existisse, teria sido encontrada –, a impossibilidade de quadrar o círculo seria uma consequência. Infelizmente, ninguém conseguiu provar que essa equação não existe. O status algébrico

de π estava no limbo. A eventual solução empregava métodos que não estavam apenas além da geometria: também encontravam-se além da álgebra.

Para apreciar a questão principal aqui, precisamos começar com uma ideia mais simples. Há uma importante distinção em matemática entre números que podem ser expressos como frações exatas $^{p}/_{q}$, onde p e q são números inteiros, e os que não podem ser assim expressos. Os primeiros são ditos racionais (são razões entre números inteiros) e os últimos são irracionais. A aproximação $^{22}/_{7}$ para π, por exemplo, é um racional. Existem aproximações melhores; uma famosa é $^{355}/_{113}$, correta até seis casas decimais. Contudo, sabe-se que nenhuma fração pode representar exatamente π: ele é irracional. Essa propriedade da qual havia muito se suspeitava foi provada pela primeira vez pelo matemático suíço Johann Heinrich Lambert em 1768. Sua prova baseia-se em uma fórmula astuciosa para função tangente em trigonometria, que ele expressou como uma fração contínua: uma quantidade infinita de frações comuns.[15] Em 1873, Charles Hermite achou uma prova mais simples baseada em fórmulas do cálculo, que foi ainda mais longe: provou que π^2 é irracional. Portanto π tampouco é raiz quadrada de um número racional.

Lambert desconfiou de algo muito mais consistente. No artigo que prova que π é irracional, ele conjecturou que π é transcendental, isto é, π não satisfaz nenhuma equação polinomial com coeficientes inteiros. Ele transcende a expressão algébrica. Descobertas subsequentes provaram que ele estava certo. O grande achado veio em duas etapas. O novo método de Hermite para provar a irracionalidade montou o cenário insinuando que o cálculo – mais precisamente sua versão rigorosa, a análise – poderia ser uma estratégia útil. Levando a ideia mais adiante, Hermite descobriu uma prova admirável de que o outro número curioso famoso em matemática, a base *e* dos logaritmos naturais, é transcendental. Em termos numéricos, *e* é aproximadamente 2,71828, e é, no mínimo, mais importante que π. A prova da transcendência de Hermite é mágica, um coelho tirado com um floreio de dentro da cartola da análise. O coelho é uma fórmula complicada relacionada com uma equação algébrica hipotética que se presume que *e* satisfaça. Usando a álgebra, Hermite prova que esta fórmula é igual a um inteiro diferente de zero. Usando a análise, ele prova que esse inteiro deve estar entre −½ e ½. Uma vez que o único inteiro nesse intervalo é zero, os resultados são contraditórios. Portanto a premissa de que *e* satisfaz uma equação algébrica deve ser falsa, portanto *e* é transcendental.

Em 1882, Ferdinand Lindemann acrescentou alguns “enfeites” ao método de Hermite, e provou que se um número diferente de zero satisfaz uma equação algébrica, então *e* elevado à potência desse número não satisfaz uma equação algébrica. Ele tirou então proveito de uma relação conhecida de Euler, envolvendo π, *e* e o número imaginário i: a famosa fórmula $e^{i\pi} = -1$. Suponha que π satisfaça alguma equação algébrica. Então o mesmo ocorre com iπ, e o teorema de Lindemann implica que −1 *não* satisfaz uma equação algébrica. No entanto, é visível que satisfaz: é a solução de $x + 1 = 0$. A única saída dessa contradição lógica é que π não satisfaça uma equação algébrica, ou seja, é transcendental. E isso significa que não se pode quadrar o círculo.

FOI UMA JORNADA LONGA e indireta desde a geometria de Euclides até a prova de Lindemann, e levou mais de 2 mil anos, mas os matemáticos finalmente chegaram lá. A história não se limita a nos dizer que o círculo não pode sofrer quadratura. É uma aula exemplar de como grandes problemas matemáticos podem ser resolvidos. Exigiu que os matemáticos formulassem cuidadosamente o que entendiam por “construção geométrica”. Tiveram de identificar características gerais de tais construções que pudessem colocar limites no que poderiam conseguir. Achar essas características exigiu fazer ligações com outra área da matemática: a álgebra. Resolver o problema algébrico, mesmo em casos simples como a construção de polígonos regulares, também envolveu a teoria dos números. Lidar com o caso difícil de π exigiu inovações adicionais, e o problema precisou ainda ser transportado para outra área da matemática: a análise.

Nenhum desses passos foi simples ou óbvio. Levou cerca de um século para completar a prova, mesmo quando as ideias principais já estavam no lugar. Os matemáticos envolvidos estavam entre os melhores de sua época, e pelo menos um estava entre os melhores de *qualquer* época. Solucionar grandes problemas requer uma compreensão profunda da matemática, mais persistência e engenhosidade. Pode envolver anos de esforço concentrado, em sua maior parte aparentemente infrutífero. Mas imagine como deve ser a sensação quando a sua persistência é recompensada e você escancara algo que frustrou a humanidade durante séculos. Como disse o presidente John F. Kennedy em 1962, ao anunciar o projeto de pouso na Lua: “Nós escolhemos … fazer [estas] … coisas, não porque são fáceis, mas porque são difíceis.”

POUCAS HISTÓRIAS EM MATEMÁTICA terminam, e π não é exceção. De tempos em tempos, surgem novas descobertas impressionantes sobre π. Em 1997, Fabrice Bellard anunciou que o trilionésimo dígito de π, em notação binária, é 1.[16] O que tornou o anúncio tão extraordinário não foi a resposta. A característica notável era que ele não calculou nenhum dos dígitos anteriores. Simplesmente tirou um dígito particular do ar.

O cálculo foi possibilitado por uma curiosa fórmula para π descoberta por David Bailey, Peter Borwein e Simon Plouffe em 1996. Ela pode parecer um pouco complicada, mas mesmo assim vamos dar uma olhada nela:

$$\pi = \sum_{n=0}^{\infty} \frac{1}{2^{4n}} \left(\frac{4}{8n+1} - \frac{2}{8n+4} - \frac{1}{8n+5} - \frac{1}{8n+6} \right)$$

O ∑ grande significa "somatória" no intervalo especificado. Aqui *n* vai de 0 a infinito (∞). Bellard na verdade usou a fórmula que ele próprio deduziu usando métodos similares, que é ligeiramente mais rápida para computações:

$$\pi = \frac{1}{64} \sum_{n=0}^{\infty} \frac{(-1)^n}{2^{10n}} \left(-\frac{32}{4n+1} - \frac{1}{4n+3} + \frac{256}{10n+1} - \frac{64}{10n+3} \right.$$
$$\left. - \frac{4}{10n+5} - \frac{4}{10n+7} + \frac{1}{10n+9} \right)$$

O ponto-chave é que muitos dos números que ocorrem aqui, 1, 4, 32, 64, 256, e também 2^{4n} e 2^{10n}, são potências de 2, que obviamente são muito simples no sistema binário usado no funcionamento interno dos computadores. Essa descoberta estimulou uma enxurrada de novas fórmulas para π e para vários outros números interessantes. O recorde para achar um único dígito binário de π é quebrado regularmente: em 2010, Nicholas Sze do Yahoo computou o dígito binário de π correspondente a 2 quatrilhões, que é 0.

A mesma fórmula pode ser usada para encontrar dígitos isolados de π em aritmética de bases 4, 8 e 16. Nada desse tipo é conhecido para

qualquer outra base; em particular, não podemos calcular dígitos decimais isoladamente. Será que tais fórmulas existem? Até a fórmula de Bailey-Borwein-Plouffe ser descoberta, ninguém imaginava que isso pudesse ser feito em binário.

---

[b] Embora o termo matematicamente correto seja “circunferência” (a linha, o contorno) e não “círculo” (a área compreendida, limitada pela circunferência), o autor usa “círculo”. Optamos por acompanhá-lo, até por ser um termo mais curto e elegante. (N.T.)

4. Mistérios na elaboração de mapas

O teorema das quatro cores

MUITOS DOS MAIORES PROBLEMAS matemáticos provêm de questões profundas e difíceis em áreas bem-estabelecidas dessa matéria. Existem grandes desafios que emergem quando uma área importante foi meticulosamente explorada. Tendem a ser bastante técnicos, e todos desse campo sabem que são difíceis de responder, porque muitos estudiosos já tentaram e falharam. A área em questão pode já dispor de muitas técnicas poderosas, pesadas máquinas matemáticas cujas manivelas podem ser giradas caso você tenha feito seu dever de casa – mas se o problema ainda permanece em aberto, então todos os meios plausíveis de utilizar essas técnicas já foram tentados, e *não funcionaram.* Logo, ou existe um meio menos plausível de usar as técnicas tentadas e testadas naquela área, ou são necessárias outras novas.

Ambas as coisas já ocorreram.

Outros grandes problemas são bem diferentes. Eles surgem do nada – traços na areia, um rabisco na margem, uma capricho passageiro. Sua formulação é simples, porque eles ainda não têm um extenso histórico matemático, não há métodos estabelecidos para se pensar sobre eles. Pode levar anos antes que a dificuldade se torne visível: pelo que se imagina, deve haver algum truque hábil, mas direto, capaz de resolver o problema em meia página. O problema das quatro cores é do segundo tipo. Levou décadas até que os matemáticos começassem a perceber quão difícil era a questão, e durante grande parte desse tempo achavam que ela *tinha sido* resolvida, em poucas páginas. Parecia ser uma questão marginal, então pouca gente se incomodou em levá-la a sério. E quando o fizeram, a solução alegada revelou-se furada. A solução definitiva tapou os "buracos", mas a essa altura a discussão havia se tornado tão complicada que exigia maciça assistência de computadores.

No longo prazo, ambos os tipos de problemas convergem, apesar de seus diferentes históricos, porque resolvê-los requer novas maneiras de pensar. Problemas do primeiro tipo podem estar embutidos numa área bem conhecida, mas os métodos tradicionais daquela área são

inadequados. Problemas do segundo tipo não pertencem a nenhuma área estabelecida – na verdade, eles motivam a criação de áreas novas –, de modo que não há métodos tradicionais que possam ser postos em ação. Em ambos os casos, solucionar o problema exige inventar novos métodos e forjar novos elos com o corpo da matemática existente.

SABEMOS EXATAMENTE de onde veio o problema das quatro cores, e não foi da matemática. Em 1852, Francis Guthrie, um jovem matemático e botânico sul-africano trabalhando para graduar-se em direito, tentava colorir os condados num mapa da Inglaterra. Ele queria assegurar-se de que quaisquer dois condados adjacentes tivessem cores diferentes, para que os limites ficassem bem claros ao simples olhar. Guthrie descobriu que precisava apenas de quatro cores distintas para executar a tarefa, e após alguma experimentação convenceu-se de que essa afirmação seria verdadeira em qualquer mapa. Por "adjacente" ele entendia que os condados em questão compartilhavam uma divisa de comprimento diferente de zero: se dois condados se tocassem num ponto, ou em diversos pontos isolados, podiam, se necessário, ter a mesma cor. Sem essa ressalva, não há limite para o número de cores, porque qualquer quantidade de regiões pode se encontrar num único ponto (Figura 8, esquerda).

Imaginando se essa formulação não seria um teorema matemático conhecido, perguntou ao seu irmão Frederick, que estudava matemática no University College, em Londres, sob orientação do distinto, mas excêntrico, Augustus De Morgan. De Morgan não sabia, então escreveu a um matemático ainda mais distinto, o irlandês sir William Rowan Hamilton:

Um aluno meu (mais tarde identificado como Frederick Guthrie) pediu-me hoje para lhe dar uma razão para um fato que eu não sabia que era fato – e ainda não sei. Ele diz que se uma figura for dividida de qualquer maneira e os compartimentos coloridos diferentemente de modo que figuras com qualquer porção de *linha* fronteiriça comum sejam coloridas de maneira diferente – quatro cores podem ser requeridas, mas não mais … A indagação é se não pode ser inventada uma necessidade para cinco ou mais … O que me diz? E isso, se for verdade, foi notado?

Frederick posteriormente referiu-se a uma "prova" que seu irmão havia sugerido, mas também disse que a ideia-chave era um desenho equivalente

à Figura 8 (direita), que prova apenas que menos que quatro cores não dá certo.

A resposta de Hamilton foi breve e não adiantou muita coisa. “É pouco provável que eu tente o seu ‘quatérnion’ de cores num futuro próximo”, escreveu ele. Na época estava trabalhando em um sistema algébrico que se tornou sua obsessão de vida, análogo aos números complexos, mas envolvendo quatro tipos de números em vez dos dois (reais e imaginários) tipos complexos. Ele os chamou de “quatérnions”. O sistema permanece importante em matemática; de fato, provavelmente é mais importante agora do que no tempo de Hamilton. Mas jamais atingiu as alturas que Hamilton almejou. Ele estava apenas fazendo uma piada acadêmica ao usar o termo, e durante muito tempo não parecia haver ligação entre os quatérnions e o problema das quatro cores. No entanto, existe uma reformulação do problema que pode ser encarada como uma afirmação sobre quatérnions, de maneira que a piada de Hamilton tem, sim, um sentido traiçoeiro.[1]

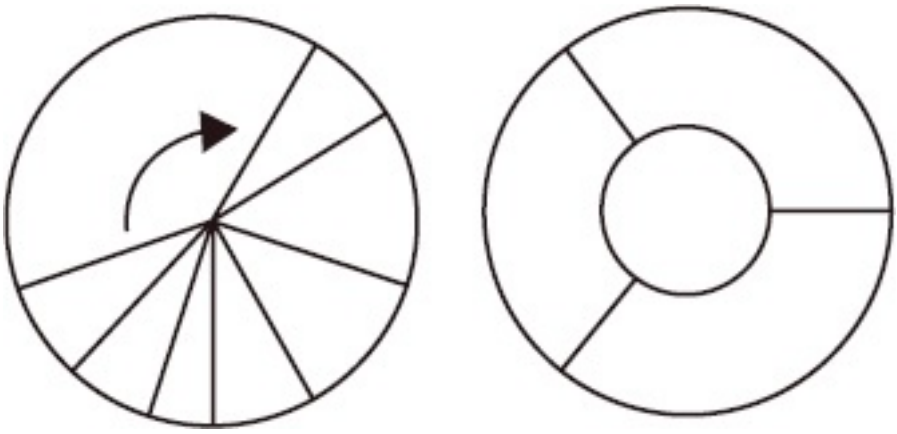


FIGURA 8 *Esquerda*: Qualquer quantidade de regiões pode se encontrar num único ponto. *Direita*: Pelo menos quatro cores são necessárias.

Incapaz de achar uma prova, De Morgan mencionou o problema aos seus conhecidos na área, na esperança de que um deles pudesse surgir com alguma ideia. No fim da década de 1860, o lógico, matemático e filósofo americano Charles Sanders Peirce alegou ter resolvido o problema das quatro cores, junto com questões semelhantes envolvendo mapas sobre superfícies mais complexas. Sua alegada prova jamais foi publicada, e é duvidoso que os métodos que ele tinha à disposição pudessem ter sido adequados.

EMBORA O PROBLEMA das quatro cores seja ostensivamente relativo a mapas, ele não tem aplicações úteis em cartografia. Critérios práticos para colorir mapas refletem sobretudo diferenças políticas, o que significa que regiões adjacentes devem ter a mesma cor, por assim dizer. O interesse do problema jaz inteiramente dentro da matemática pura, numa área nova que recém-começara a se desenvolver: a topologia. Essa é a "geometria da folha de borracha", na qual as formas podem ser deformadas de qualquer maneira contínua. Mas mesmo aí o problema das quatro cores não pertencia à vertente principal. Parecia não passar de uma curiosidade menor.

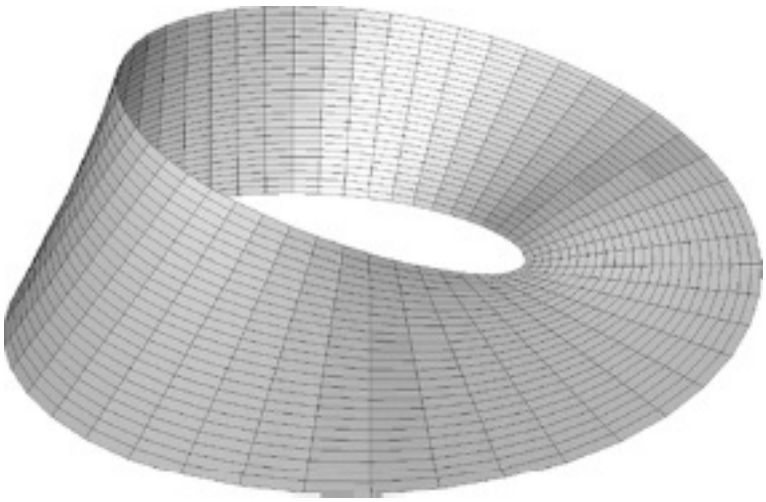


FIGURA 9 A faixa de Möbius tem só um lado.

Um dos pioneiros da topologia foi August Möbius, hoje famoso pela sua faixa de um lado só (Figura 9). Você pode fazer uma dessas pegando uma tira de papel, girando uma das pontas a 180 graus e grudando as extremidades de modo a formar um anel. Um de seus amigos, o linguista Benjamin Weiske, apresentou a Möbius um quebra-cabeça: poderia um rei indiano com cinco filhos, todos príncipes, dividir seu reino de maneira que a região pertencente a cada príncipe compartilhasse uma fronteira de comprimento diferente de zero com a região pertencente aos outros quatro príncipes? Möbius repassou o quebra-cabeça para seus alunos como exercício. Mas na aula seguinte, desculpou-se por pedir-lhes que realizassem o impossível. Com isso ele quis dizer que podia *provar* que era impossível.[2]

É difícil atacar esse quebra-cabeça geometricamente, porque as formas das regiões e a maneira como estão arranjadas pode ser em princípio muito complicada. O progresso depende de uma grande simplificação: tudo que

realmente importa é quais regiões são adjacentes a outras, e como as fronteiras comuns estão dispostas umas em relação às outras. Isso é informação topológica, independente das formas precisas. E pode ser representada de maneira clara e simples, num diagrama – ou, como se diz hoje, em um grafo, que é um termo mais evocativo.

Um grafo é um conceito arrasadoramente simples: um conjunto de vértices, representados por pontos, alguns dos quais ligados por margens, desenhadas como traços. Pegue qualquer mapa, por exemplo a Figura 10 (esquerda). Para convertê-lo em um grafo, coloque um ponto dentro de cada região (Figura 10, centro). Sempre que duas regiões tiverem uma fronteira extensa, desenhe um traço unindo os pontos correspondentes, cruzando essa fronteira. Se houver diversos segmentos de fronteira comum separados, cada um recebe seu próprio traço. Faça isso para todas as regiões e para todos os trechos de fronteira, de modo que os traços não se cruzem, nem sobre si mesmos, e juntem-se apenas nos pontos. Então jogue fora o mapa original e guarde apenas pontos e traços. Estes formam o grafo dual do mapa (Figura 10, direita).

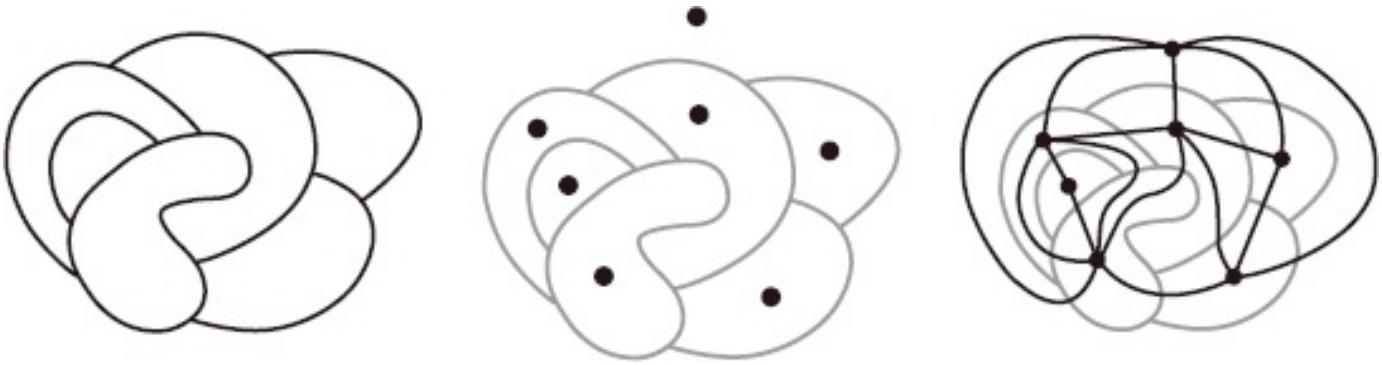


FIGURA 10 *Esquerda*: Um mapa. *Centro*: Coloque um ponto em cada região. *Direita*: Ligue os pontos cruzando as fronteiras para formar o grafo dual (apenas pontos e traços pretos).

A palavra “dual” é usada porque o procedimento toma regiões, traços e pontos (junções entre regiões do mapa) e os transforma em pontos, traços e regiões. Uma região no mapa corresponde a um ponto no grafo dual. Um segmento de fronteira no mapa corresponde a um traço no grafo dual; não o mesmo traço, mas outro que cruza a fronteira ligando os pontos correspondentes. Um ponto no mapa, onde três ou mais regiões se encontram, corresponde a uma região no grafo dual, delimitada por uma alça de traços fechada. Assim, o grafo dual é em si um mapa, porque os traços delimitam regiões, fazendo o dual do dual ser o mapa original, desconsiderando alguns aspectos técnicos que excluem pontos e traços

desnecessários.

O problema dos cinco príncipes pode ser reinterpretado usando o grafo dual: é possível ligar cinco pontos do plano por traços, sem cruzamentos? A resposta é "não", e a chave é a fórmula de Euler, que afirma que se um mapa no plano consiste de F faces (regiões), A arestas (traços) e V vértices (pontos), então $F + V - A = 2$. Aqui contamos o resto do plano, fora do grafo, como uma grande região. Essa fórmula foi um dos primeiros indícios de que valia a pena investigar considerações topológicas, e ela reaparece no Capítulo 10.

A prova de que o quebra-cabeça dos príncipes indianos é impossível inicia assumindo que existe uma solução, e deduz uma contradição. Qualquer solução terá $V = 5$, o número de pontos. Como cada par de pontos é ligado por um traço, e há dez pares, $A = 10$. O teorema de Euler implica que $F = A - V + 2 = 7$. As regiões do grafo dual estão cercadas por alças de traços fechadas, e apenas um traço liga qualquer par de pontos; portanto essas alças devem conter pelo menos três traços. Como há sete regiões, isso produz pelo menos 21 traços… exceto que cada traço esteja sendo contado duas vezes porque separa duas regiões. Então há pelo menos dez traços e meio. O número de traços precisa ser inteiro, então na verdade deve haver pelos menos onze traços. No entanto, já sabemos que há dez traços. Isso é uma contradição lógica, e prova que tal grafo não existe. O rei não pode dividir sua terra da maneira prescrita.

O aspecto encorajador desse argumento é que refinados métodos topológicos nos possibilitam provar algo específico e interessante a respeito dos mapas. Contudo, ao contrário da errônea concepção comum, que De Morgan parece ter partilhado, a impossibilidade de resolver o quebra-cabeça dos cinco príncipes indianos *não* prova o teorema das quatro cores. Uma prova pode estar errada mesmo que a conclusão esteja correta, ou que se ignore estar incorreta. Se em algum ponto em uma prova alegada eu encontro um triângulo com quatro lados, posso parar de ler, porque a prova está errada. Não importa o que aconteça depois disso ou qual seja a conclusão. A nossa resposta para o quebra-cabeça dos príncipes indianos mostra que um modo específico de refutar o teorema das quatro cores não funciona. Todavia, isso não quer dizer que nenhum *outro* modo possa funcionar. Potencialmente, poderia haver muitos obstáculos para as quatro cores (de agora em diante usarei esse termo em vez do desconfortável "colorir o mapa com quatro cores"). A existência de cinco regiões, todas adjacentes entre si, é meramente um desses obstáculos. Pelo

que sabemos, poderia haver um mapa muito complicado com 703 regiões, de tal maneira que, não importa como você pinte com as quatro cores 702 delas, a última região sempre necessitará de uma quinta cor. Essa região teria de se encostar em pelo menos outras quatro, mas isso é inteiramente factível, e não requer um arranjo do tipo dos príncipes indianos. Se existisse um mapa como esse, ele provaria que quatro cores não bastam. Qualquer prova precisa excluir esse tipo de obstáculo. E essa afirmação é válida mesmo que eu não lhe mostre – não possa lhe mostrar – um exemplo explícito de tal obstáculo.

DURANTE ALGUM TEMPO o problema das quatro cores parecia ter desaparecido sem deixar vestígios, mas voltou à tona em 1878 quando Arthur Cayley o mencionou em um encontro da London Mathematical Society. Apesar do nome, essa organização representava o conjunto da matemática britânica (ou pelo menos inglesa), e seu fundador foi De Morgan. Cayley perguntou se alguém tinha conseguido alguma solução. Sua indagação foi publicada logo em seguida na revista científica *Nature*. Um ano depois ele escreveu um artigo mais extenso para a *Proceedings of the Royal Geographical Society*.[3] Parecia ser o lugar lógico para tal artigo, porque o problema trata ostensivamente de mapas. Pode até ser que ele tenha sido solicitado a apresentar o artigo. Mas na verdade não foi uma escolha sensata, porque nenhum elaborador de mapas teria qualquer motivo para querer saber a resposta, além da mera curiosidade. Infelizmente, a escolha dessa revista significou que poucos matemáticos estariam cientes da existência do artigo. O que foi uma pena, porque Cayley explicou o motivo pelo qual o problema podia ser traiçoeiro.

No Capítulo 1, eu disse que uma prova é um pouco como uma batalha. O militar reconhece a diferença entre tática e estratégia. A tática é como você vence conflitos locais; a estratégia estabelece a estrutura ampla da campanha. A tática envolve movimentos de tropa detalhados; a estratégia envolve planos amplos, com margem para muitas decisões táticas diferentes em qualquer etapa. O artigo de Cayley concentrava-se em táticas, mas continha um vago indício de uma estratégia que, com o correr do tempo, acabou derrubando o problema das quatro cores. Ele observou que somar regiões uma de cada vez não funcionava se fosse seguida a linha de raciocínio óbvia. Mas talvez funcionasse caso se adotasse uma linha de raciocínio menos óbvia.

Suponha que você pegue um mapa, remova uma região – digamos, fundindo-a com uma outra vizinha, ou encolhendo-a até transformá-la em um ponto. Suponha que o mapa resultante possa ser de quatro cores. Agora coloque de volta a região original. Se tiver sorte, suas vizinhas talvez usem apenas três cores. Então tudo que você precisa fazer é colori-la com a quarta cor. O argumento de Cayley era que esse procedimento poderia não dar certo, porque os vizinhos da última região poderiam usar quatro cores distintas. Mas isso não significa que você esteja encalhado. Há duas maneiras de se contornar o obstáculo: você pode ter escolhido a região errada, ou o modo errado de colorir o mapa menor.

Ainda com base em suposições não substanciais (esse é um meio muito eficaz de se ter ideias de pesquisa, embora em algum momento você tenha que substanciá-las), vamos admitir que algo desse tipo sempre possa ser estabelecido. Então, isso lhe diz que um mapa pode ter quatro cores se um mapa menor tem quatro cores. Pode não parecer um grande progresso: como sabemos se o mapa menor pode ter quatro cores? A resposta é que o mesmo procedimento aplica-se a mapas menores, levando a um mapa ainda menor… e assim por diante. No final, você obtém um mapa tão pequeno que possui apenas quatro regiões, e aí você sabe que pode ter quatro cores. Agora, inverta a sequência, colorindo mapas ligeiramente maiores em cada etapa… e, por fim, volte ao seu mapa original.

Essa linha de raciocínio é chamada de "prova por indução matemática". É um método padrão com uma formulação mais técnica, e a lógica por trás dele pode ser rigorosa. A estratégia de prova proposta por Cayley torna-se mais transparente se o método for reformulado usando um conceito logicamente equivalente: a do contraexemplo mínimo. Nesse contexto, um contraexemplo é qualquer mapa hipotético que não pode ter quatro cores. Tal mapa será mínimo se qualquer mapa com uma quantidade menor de países *puder* ser colorido com quatro cores. Se existe o contraexemplo, ele deve ser mínimo: basta escolher um contraexemplo com o menor número possível de regiões. Portanto, se contraexemplos mínimos não existirem, então não existirão contraexemplos. E se não existem contraexemplos, o teorema das quatro cores deve ser verdadeiro.

O procedimento de indução se resume a isso. Suponha que possamos provar que sempre é possível colorir um contraexemplo mínimo com quatro cores, na condição de que algum mapa menor correlacionado possa ser assim colorido. Então o contraexemplo mínimo não pode ser realmente contraexemplo. Justamente porque o mapa é mínimo, *todos* os mapas

menores podem ter quatro cores, então pelo que supusemos que possa ser provado, o mesmo vale para o mapa original. Portanto não há contraexemplos mínimos, então não há contraexemplos. Essa ideia desvia o foco do problema de todos os mapas para apenas os contraexemplos mínimos hipotéticos, e especificando um procedimento de redução – um meio sistemático de transformar um mapa menor de quatro cores correlato em um mapa original de quatro cores.

Por que ficar se metendo com contraexemplos mínimos, em vez de simples contraexemplos? É uma questão de técnica. Mesmo que a princípio não saibamos se contraexemplos existem, uma das características paradoxais, porém úteis, dessa estratégia é que podemos dizer muita coisa sobre como seria a aparência dos mínimos se eles existissem.

Isso requer a capacidade de pensar logicamente sobre aspectos hipotéticos, uma habilidade vital para qualquer matemático. Para oferecer um sabor do processo, vou provar o teorema das *seis* cores. Para fazê-lo, tomamos emprestado um truque do quebra-cabeça dos cinco príncipes, e reformulamos tudo em termos de grafo dual, no qual as regiões tornam-se pontos. O problema das quatro cores é então equivalente a outra questão: dado um grafo no plano cujos traços não se cruzam, é possível colorir os *pontos* com quatro cores de maneira que dois pontos interligados por um traço sempre tenham cores diferentes? A mesma reformulação aplica-se a qualquer número de cores.

Para ilustrar o poder dos contraexemplos mínimos, vou usá-los para provar que qualquer grafo planar pode ser colorido com seis cores. Mais uma vez, a principal ferramenta técnica é a fórmula de Euler. Dado um ponto num grafo dual, definimos como seus vizinhos aqueles que estão ligados a ele por um traço. Um ponto pode ter muitos vizinhos, ou apenas alguns poucos. Pode-se demonstrar que a fórmula de Euler implica que alguns pontos devem ter poucos vizinhos. Mais precisamente, num grafo planar é impossível que todos os pontos tenham seis ou mais vizinhos. Inseri nas notas uma prova para esse fato, para evitar interromper o fluxo de ideias.[4] Isso fornece a alavanca necessária para começar a levantar o problema em partes. Considere um contraexemplo mínimo hipotético para o teorema das seis cores. Ele é um grafo que não pode receber seis cores. Agora provo que esse mapa não pode existir. Como consequência da fórmula de Euler citada, ele contém pelo menos um ponto com cinco vizinhos ou menos. Vamos apagar temporariamente esse ponto e os traços que o ligam aos vizinhos. O grafo resultante tem menos pontos, então por

minimalidade pode ter seis cores. (É aqui que encalhamos, a menos que o nosso contraexemplo hipotético seja mínimo.) Agora ponha de volta o ponto e os traços apagados. Esse ponto tem no máximo cinco vizinhos, então há sempre uma sexta cor. Use essa cor no ponto apagado. Agora conseguimos colorir com seis cores nosso contraexemplo mínimo – mas isso contradiz sua minimalidade. Então não existem contraexemplos mínimos para o teorema das seis cores, e isso significa que o teorema das seis cores é verdadeiro.

Isso é encorajador. Até agora, pelo que sabíamos, alguns mapas poderiam precisar de vinte cores, ou 703, ou milhões. Agora sabemos que mapas como esse não são mais reais do que um pote de ouro no fim do arco-íris. Um número específico, limitado, de cores serve para *qualquer* mapa. Esse é um triunfo genuíno para contraexemplos mínimos, e encorajou os matemáticos a ajustar o argumento na esperança de substituir seis cores por cinco, ou, se você fosse realmente esperto, quatro.

TODOS OS CRIMINOSOS necessitam de advogados. Um causídico chamado Alfred Kempe estava presente ao encontro no qual Cayley mencionou o problema das quatro cores. Ele estudara matemática com Cayley como aluno de graduação em Cambridge, e seu interesse pelo assunto não diminuiu. No período de um ano, Kempe convenceu-se de que havia resolvido o problema, e publicou sua solução em 1879 no recém-fundado *American Journal of Mathematics*. Um ano depois, publicou uma prova simplificada, que corrigia alguns erros da primeira. Ressaltava que

Uma alteração muito pequena em uma parte do mapa pode tornar necessário recolori-lo totalmente. Após uma busca um tanto árdua, eu consegui … acertar o ponto fraco, que se mostrou fácil de atacar.

Vou reinterpretar as ideias de Kempe em termos de grafo dual. Mais uma vez, ele começou a partir da fórmula de Euler e a consequente existência de um ponto com três, quatro ou cinco vizinhos. (Um ponto com dois vizinhos jaz no meio do traço, e não contribui em nada para o grafo ou o mapa, podendo ser tranquilamente omitido.)

Se existe um ponto com três vizinhos, o procedimento que usei para provar o teorema das seis cores aplica-se quando há apenas quatro cores. Remova o ponto e os traços que nele se cruzam, ponha quatro cores no

restante, coloque o ponto e os traços de volta, use uma cor reserva para o ponto. Podemos, portanto, assumir que nenhum ponto tem três vizinhos.

Se existe um ponto com quatro vizinhos a tática acima falha, porque uma cor reserva pode não estar disponível. Kempe divisou um método sagaz de lidar com esse obstáculo: de qualquer modo, apague aquele ponto, mas depois disso mude as cores do mapa menor resultante de maneira que dois desses quatro vizinhos tenham a mesma cor. Após essa mudança, os vizinhos do ponto apagado usam no máximo quatro cores, deixando uma cor reserva para o ponto apagado. A ideia básica de Kempe de recolorir o esquema é que dois dos pontos vizinhos devem ter cores diferentes – digamos vermelho e azul, com as outras cores sendo verde ou amarelo. Se ambos forem verdes ou ambos amarelos, a outra cor estará disponível para o ponto apagado. Assim, podemos assumir que um é verde e um, amarelo. Agora encontre todos os pontos que podem ser ligados ao azul por uma sequência de traços, usando apenas pontos azuis e vermelhos. Vamos chamar essa sequência de corrente azul-vermelho de Kempe.[5] Por definição, todo vizinho de qualquer ponto na corrente de Kempe, que não esteja ele próprio participando da corrente de Kempe, é ou verde ou amarelo, porque um vizinho azul ou vermelho já estará na corrente. Tendo descoberto tal corrente, observe que trocar a cor azul ou vermelha para todos os pontos da corrente produz outro colorido do grafo, ainda satisfazendo a condição-chave de que pontos adjacentes tenham cores diferentes (Figura 11).

Se o vizinho vermelho do nosso ponto original não está nessa corrente azul-vermelha, faça essa mudança. O vizinho azul dos pontos originais vira vermelho; o vizinho vermelho permanece vermelho. Agora os vizinhos do ponto original usam no máximo três cores diferentes: vermelho, verde e amarelo. Isso deixa o azul para o ponto original, e pronto. No entanto, a corrente azul-vermelha poderia dar uma volta e juntar-se ao vizinho azul. Se isso acontecer, deixe a corrente azul e vermelha em paz, e use o mesmo truque para os vizinhos amarelo e verde do ponto original. Comece com o verde e forme uma corrente de Kempe verde-amarela. Essa corrente *não pode* se juntar com o vizinho amarelo, porque a corrente anterior azul-vermelha está no caminho. Troque amarelo e verde, e pronto.

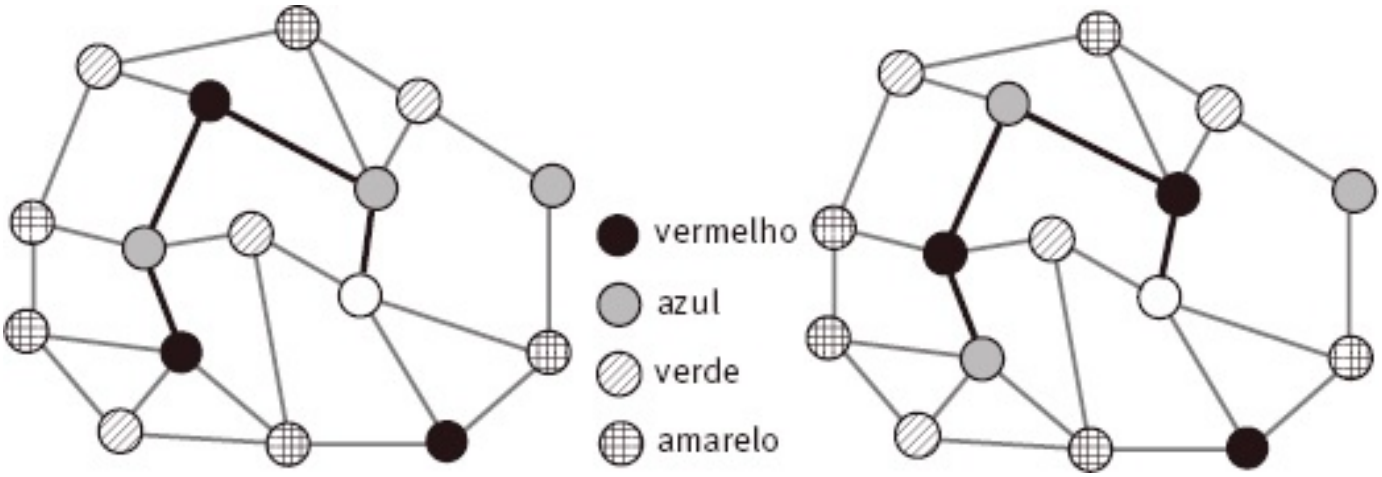


FIGURA 11 Troca de cores numa corrente de Kempe (traços pretos fortes) associada a um ponto de grau 4 (branco) que tem vizinhos de todas as quatro cores. *Esquerda*: Cores originais. *Direita*: Com cores trocadas, o azul fica disponível para o ponto branco.

Isso permite um último caso, quando não há pontos com três ou quatro vizinhos, mas pelo menos um tem cinco vizinhos. Kempe propôs uma regra similar, porém mais complicada, de recolorir, que parecia solucionar também esse caso. Conclusão: o teorema das quatro cores é verdadeiro, e Kempe o tinha provado. E chegou inclusive até a mídia: *The Nation*, uma revista norte-americana, mencionou a solução em sua seção de resenhas.

A prova de Kempe parecia ter dado descanso ao problema. Para a maioria dos matemáticos era fato consumado. Peter Guthrie Tait continuou publicando artigos sobre o problema, buscando uma prova mais simples; isso o levou a algumas descobertas proveitosas, mas a prova mais simples lhe escapou.

ENTRA EM CENA Percy Heawood, um professor de matemática na Universidade de Durham conhecido pelo apelido de "Pussy", graças ao seu magnífico bigode. Como aluno de graduação em Oxford, aprendera o problema das quatro cores com Henry Smith, catedrático de geometria. Smith disse-lhe que o teorema, embora provavelmente verdadeiro, não era provado, e então Heawood tinha uma chance. Ao longo do caminho deparou-se com o artigo de Kempe, e tentou entendê-lo. Publicou o resultado em 1889 como "Teorema das cores do mapa", lamentando que o objetivo do artigo fosse mais "destrutivo do que construtivo, pois será demonstrado que há um defeito na atualmente reconhecida prova". Kempe cometera um erro.

Era um erro sutil, e ocorria no método de recolorir quando o ponto

apagado tinha cinco vizinhos. O esquema de Kempe podia ocasionalmente mudar a cor de algum ponto como efeito retroativo de mudanças posteriores. Mas Kempe admitira que uma vez mudada a cor de um ponto, ela não voltaria a mudar. Heawood encontrou um grafo para o qual o esquema de recolorir de Kempe dava errado, então sua prova era falha. Kempe foi rápido em reconhecer o erro, e acrescentou que "não conseguira remediar o defeito". O teorema das quatro cores estava novamente à disposição.

Heawood extraiu do desastre algumas migalhas de consolo para Kempe: seu método era bem-sucedido em provar o teorema das cinco cores. Heawood também trabalhou em duas generalizações do problema: impérios, em que cada região pode consistir de várias partes desconectadas, todas exigindo a mesma cor, e mapas sobre superfícies mais complicadas. A questão análoga numa esfera tem a mesma resposta que a do plano. Imagine um mapa numa esfera, e gire-a até que o polo norte esteja em algum ponto dentro de uma região. Se você apagar o polo norte, poderá abrir a esfera pontuada para obter um espaço que é topologicamente equivalente a um plano infinito. A região que contém o polo torna-se a infinitamente grande, cercando o resto do mapa. Mas há outras superfícies mais interessantes. Entre elas está o toro, que tem o formato de uma rosquinha com um furo (Figura 12, esquerda).

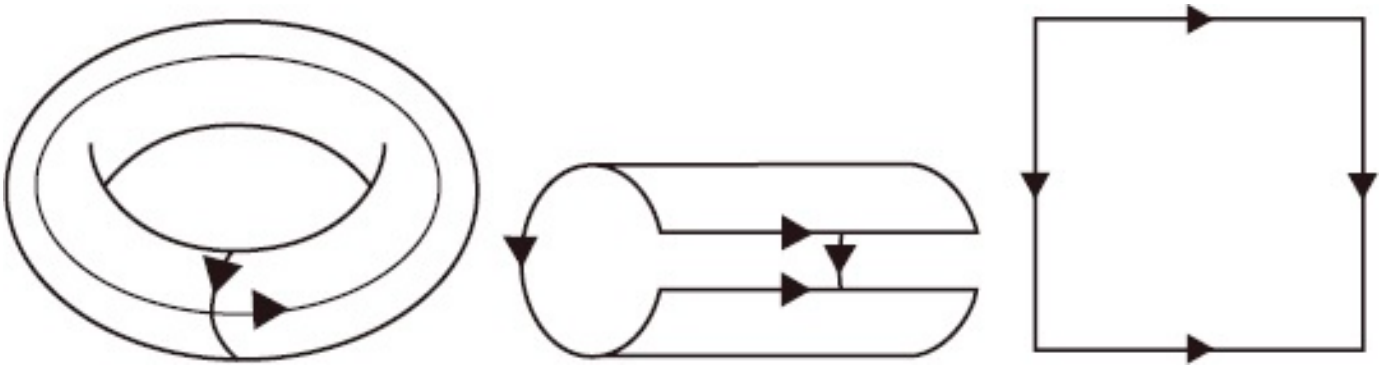


FIGURA 12 Cortar um toro, abri-lo e desenrolá-lo de modo a formar um quadrado.

Existe um jeito útil de visualizar o toro, que costuma simplificar a vida. Se cortarmos o toro seguindo duas curvas fechadas (Figura 12, centro), podemos abri-lo e transformá-lo num quadrado (Figura 12, direita). Essa transformação altera a topologia do toro, mas podemos contornar a situação "igualando" bordas opostas do quadrado. Com efeito (e uma definição rigorosa torna essa ideia precisa), concordamos em tratar pontos

correspondentes dessas bordas como se fossem idênticos. Para ver como funciona, inverta a sequência das figuras. O quadrado é enrolado, e bordas opostas realmente se grudam. Agora vem a parte inteligente: você não precisa efetivamente enrolar o quadrado e fazer coincidir as bordas correspondentes. Basta trabalhar com o quadrado plano, contanto que tenha em mente a regra de "igualar" as bordas. Tudo que você faz no toro, como, por exemplo, desenhar figuras sobre ele, tem uma construção precisa correspondente no quadrado.

Heawood provou que sete cores são igualmente necessárias e suficientes para colorir qualquer mapa sobre um toro. A Figura 13 (esquerda) mostra que sete são necessárias, usando o quadrado para representar o toro, como acabamos de descrever. Observe como as regiões encaixam-se nas bordas opostas. Há superfícies como o toro, porém com mais furos (Figura 13, direita). O número de furos é chamado genus, sendo representado pela letra g. Heawood conjecturou uma fórmula para a quantidade de cores exigida num toro com g furos, quando $g \geq 1$: é o menor número inteiro inferior ou igual a

$$\frac{7 + \sqrt{48g + 1}}{2}$$

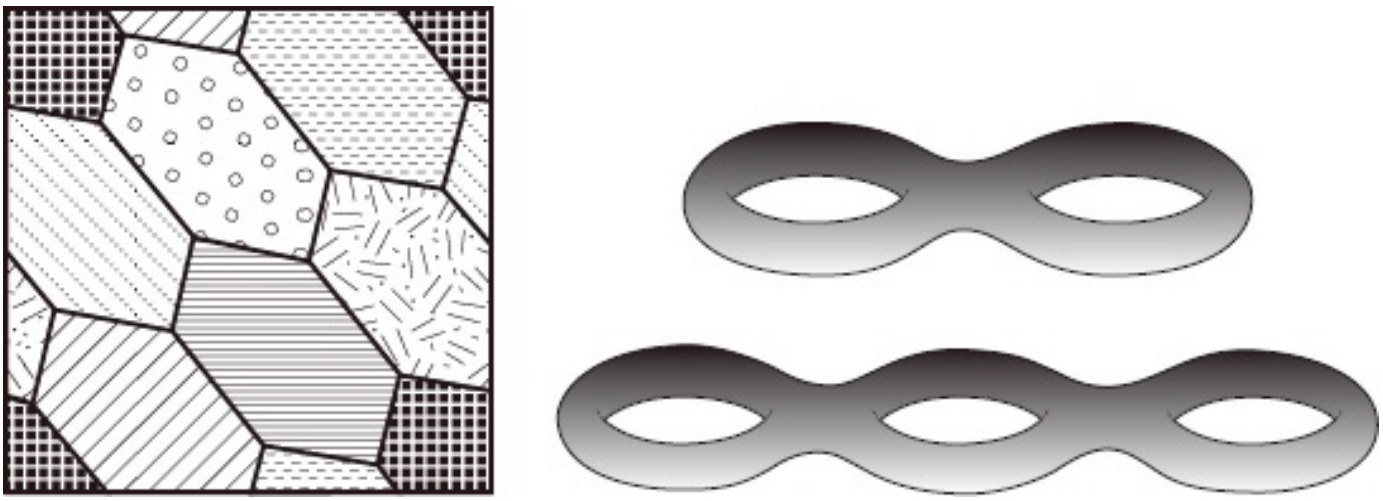


FIGURA 13 *Esquerda:* Mapa de sete cores num toro. O toro é representado como um quadrado cujos lados opostos estão conceitualmente "dando a volta" de modo que se grudam. As regiões do mapa devem se ajustar manualmente ao longo das bordas correspondentes. *Direita:* Toros com dois e três furos.

Quando g varia de 1 a 10, essa fórmula nos dá os números

7 8 9 10 11 12 12 13 13 14

A quantidade de cores especificadas pela fórmula cresce mais lentamente que o genus, e muitas vezes não faz diferença se você inclui um furo extra no toro. Isso é uma surpresa, porque cada furo extra proporciona mais liberdade para inventar mapas mais complicados.

Heawood não tirou essa fórmula do ar sem mais nem menos. Ela surgiu ao se generalizar a maneira como provei o teorema das seis cores no plano. Ele foi capaz de provar que esse número de cores é sempre suficiente. A grande questão, porém, era se esse número podia ser diminuído. Exemplos de valores pequenos do genus sugeriam que a estimativa de Heawood era a melhor possível. Em 1968, após uma prolongada investigação, Gerhard Ringel e John W.T. (Ted) Youngs preencheram os detalhes finais em uma prova, construída em seu próprio trabalho e no de vários outros, demonstrando que isso é correto. Seus métodos são combinatórios, baseados em tipos especiais de grafos, e suficientemente complicados para encher um livro inteiro.[6]

QUANDO $g = 0$, ou seja, para mapas sobre uma esfera, a fórmula de Heawood nos dá quatro cores, mas sua prova de suficiência não funciona na esfera. Apesar do impressionante progresso para superfícies com ao menos um furo, o problema original das quatro cores ainda estava para ser resolvido. Os poucos matemáticos dispostos a dedicar esforços sérios à questão conformaram-se com aquilo que, em termos de uma guerra, provavelmente seria um cerco prolongado. O problema era um castelo fortemente defendido; eles tinham esperança de construir máquinas de assédio cada vez mais poderosas e ir derrubando partes até caírem totalmente os muros do castelo. E as máquinas eram construídas, e os muros não caíam. No entanto, os atacantes foram aos poucos acumulando um bocado de informação sobre como não resolver o problema, e os tipos de obstáculo que pareciam inevitáveis. A partir desses fracassos, começou a emergir uma estratégia ambiciosa. Era uma extensão natural dos métodos de Kempe e Heawood, e veio em três partes. Vou mencioná-las usando o grafo dual, o ponto de vista padrão dos dias de hoje:

1. Considere um contraexemplo mínimo.
2. Encontre uma lista de configurações inevitáveis: grafos menores, com a propriedade de que qualquer contraexemplo mínimo deve conter algo na lista.

3. Prove que cada uma das configurações inevitáveis é redutível. Ou seja: se um grafo menor, obtido deletando-se a configuração inevitável, pode ter quatro cores, então essas cores podem ser redistribuídas de modo tal que quando a configuração inevitável for restaurada, o colorido de quatro cores do grafo menor se estenda a todo o grafo.

Juntando esses três passos, podemos provar que um contraexemplo mínimo não existe. Se existisse, conteria uma configuração inevitável. Mas o resto do grafo é menor, então seu caráter mínimo implica que pode ter quatro cores. A redutibilidade agora implica que o grafo original pode ter quatro cores. E isso é uma contradição.

Nesses termos, Kempe descobrira corretamente uma lista de configurações inevitáveis: um ponto com três linhas saindo, um com quatro e um com cinco (Figura 14). E também provou corretamente que as duas primeiras são redutíveis. Seu erro residiu na prova de que a terceira configuração é redutível. Ela não é. Proposta: substitua essa configuração ruim por uma lista mais longa, assegurando-se de que a lista permaneça inevitável. Faça isso de tal maneira que cada configuração da nova lista seja redutível. Ou seja: procure uma lista inevitável de configurações redutíveis. Se conseguir, terá provado o teorema das quatro cores.

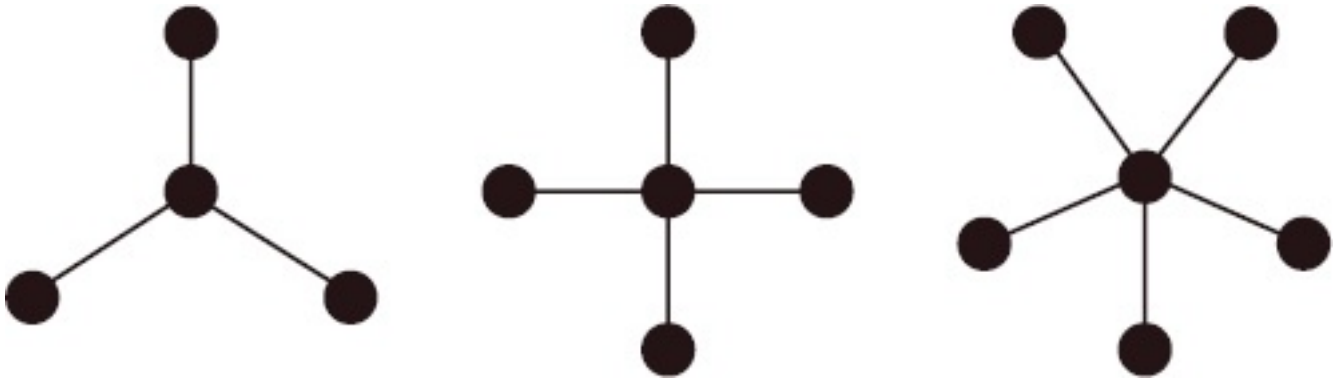


FIGURA 14 Lista de configurações inevitáveis de Kempe.

Pode ser que não haja tal lista, mas essa estratégia vale a pena ser tentada, e ninguém teve uma ideia melhor. Contudo, ela tem uma desconfortável tensão interna. De um lado, quanto mais longa a lista, maior a chance de ser inevitável, o que é bom. De outro lado, quanto mais longa a lista, menos provável que cada configuração nela seja redutível. Se uma única configuração não for redutível, então toda a prova cai por terra, e

esse perigo torna-se mais agudo à medida que a lista cresce. O que é ruim. Por um *terceiro* lado… uma lista mais longa oferece mais oportunidades de escolher configurações redutíveis, o que é bom. Por um *quarto* lado, aumenta o trabalho necessário para provar a redutibilidade, o que é ruim. E por um *quinto* lado, de qualquer modo, não havia bons métodos para se fazer isso, o que era pior.

Esse tipo de coisa é o que faz um grande problema ser grande.

Assim, por algum tempo, pequenos pedaços do castelo foram sendo ocasionalmente derrubados, e sua queda não fazia a menor diferença para a solidez da fortaleza; entrementes, a corrente principal da matemática bocejava, se é que era notada. Mas alguém estava construindo uma catapulta melhor, e seu nome era Heinrich Heesch. Sua grande contribuição foi um modo sistemático de provar que uma configuração é redutível. Ele o chamou de "descarga", e era mais ou menos análogo a imaginar que os pontos do grafo carregam cargas elétricas, e permitir que a eletricidade flua de um ponto a outro.

Mesmo com esse método, achar à mão um conjunto inevitável de configurações redutível seria uma tarefa assustadora. As configurações individuais provavelmente seriam bastante pequenas, mas teriam de ser muitas. Heesch perseverou, e em 1948 ministrou um curso onde sugeriu que seriam necessárias cerca de 10 mil configurações. A essa altura, ele já provara que quinhentas candidatas eram redutíveis. Na plateia estava um jovem chamado Wolfgang Haken, que mais tarde disse não ter entendido muita coisa das aulas de Heesch, mas alguns dos pontos principais tinham ficado na sua cabeça. Haken passou a estudar topologia, fazendo uma importante descoberta na teoria dos nós. Isso o encorajou a trabalhar na conjectura de Poincaré (Capítulo 10). Para uma linha de ataque específica, ele classificou as possibilidades em duzentos casos, solucionou 198 deles e debateu-se com os dois restantes por treze anos. Nesse ponto, ele desistiu, e em vez disso começou a trabalhar com o problema das quatro cores. Haken claramente gostava de problemas difíceis, mas estava preocupado com a possibilidade de acontecer algo semelhante nas 10 mil configurações de Heesch. Imagine ter êxito com 9.998 e ficar encalhado com as duas últimas. Assim, em 1967, convidou Heesch para visitar a Universidade de Illinois, onde ficava sua base, para pedir seu conselho.

Naquele tempo os computadores estavam começando a ser úteis para a matemática real, mas eram máquinas enormes localizadas em algum edifício central, não coisas que cabiam em cima da escrivaninha ou dentro

da maleta. Haken imaginou se eles poderiam ajudar. Heesch já tivera a mesma ideia e fez uma estimativa aproximada da complexidade do problema. Essa estimativa indicou que o melhor computador disponível para ele não estava à altura da tarefa. Illinois tinha um supercomputador mais potente, o ILLIAC-IV, de modo que Haken requisitou tempo para utilizá-lo. Mas o supercomputador não estava pronto, então disseram-lhe para tentar o Cray 6600 no Laboratório Brookhaven, em Long Island. O diretor do centro de computadores do laboratório era Yoshio Shimamoto, que havia muito era fascinado pelo problema das quatro cores – um golpe de sorte que deu a Heesch e Haken acesso à máquina.

O COMPUTADOR CORRESPONDEU às expectativas, mas Haken começou a se perguntar se poderia ser usado com mais eficiência. Eles estavam gerando montes de configurações redutíveis na esperança de juntar uma lista inevitável, mas a estratégia desperdiçava muito tempo em configurações potenciais que acabavam revelando-se não redutíveis. Por que não percorrer o caminho inverso: tornar a inevitabilidade o objetivo principal e checar a redutibilidade depois? É claro que era preciso usar configurações que tivessem uma boa chance de ser redutíveis, mas parecia um caminho melhor. Àquela altura, porém, o Cray em Brookhaven estava sendo usado para coisas mais importantes. Pior, vários entendidos disseram a Haken que os métodos que ele pretendia usar não podiam ser, de jeito nenhum, transformados em programas de computador. Ele acreditou e proferiu uma palestra dizendo que o problema não podia ser resolvido sem computadores, mas agora parecia que tampouco podia ser resolvido com os aparelhos. Decidira desistir.

Na plateia estava um perito em programação, Kenneth Appel, que disse a Haken que os alegados entendidos estavam simplesmente tentando deixá-lo de fora porque os programas exigiriam muito trabalho e o resultado era muito incerto. Na opinião de Appel, não havia problema matemático que não pudesse ser programado. A questão crucial era se o programa chegaria ao objetivo dentro de um tempo razoável. Ambos juntaram forças. A estratégia evoluiu à medida que aperfeiçoamentos no método da descarga foram provocando mudanças no programa, e aperfeiçoamentos do programa levaram a mudanças no método da descarga. Isso os conduziu a um novo conceito: configurações "geograficamente boas", que não continham certas configurações desagradáveis que impediam a redutibilidade. A chance de tais

configurações serem redutíveis foi bastante melhorada, e a propriedade definidora era fácil de verificar. Appel e Haken decidiram provar teoricamente, em vez de usar o computador, que havia uma lista inevitável de configurações geograficamente boas. Em 1974, tiveram sucesso.

Isso era encorajador, mas eles sabiam o que era passível de acontecer. Algumas de suas configurações geograficamente boas revelar-se-iam não redutíveis, e eles teriam de removê-las e substituí-las por uma lista mais longa e complicada. Os cálculos estariam perseguindo o próprio rabo, e dariam certo só se conseguissem apanhá-lo. Em vez de perder anos numa busca infrutífera, fizeram alguns cálculos grosseiros para avaliar quanto tempo o processo poderia levar. Os resultados foram levemente animadores, então o trabalho continuou. Teoria e computação alimentavam-se e modificavam-se mutuamente. Às vezes, o computador parecia ter mente própria, "descobrindo" características úteis de configurações. Então, a administração da universidade comprou para seu uso um computador novo, muito potente – mais potente do que os disponíveis para os cientistas da universidade. Após algum protesto, e questões contundentes, metade do tempo da máquina foi disponibilizada para uso científico. A lista sempre mutante de configurações inevitáveis examinada por Appel e Haken estabilizou-se em torno de 2 mil delas. Em junho de 1976, o computador expeliu o resultado de sua última verificação de redutibilidade, e a prova estava completa. A história chegou à mídia, começando no *Times* e espalhando-se rapidamente ao redor do mundo.

Eles ainda precisavam ter certeza de que não havia erros bobos, e a essa altura várias outras equipes estavam nessa busca acalorada. Em julho, Appel e Haken estavam confiantes no funcionamento de seu método, e anunciaram sua prova oficialmente para a comunidade matemática fazendo circular um pré-impresso – uma cópia barata do rascunho de um artigo, com intenção de publicação posterior. Na época, era típico levar de um a dois anos para pôr em versão impressa uma prova matemática. Para evitar retardar o progresso, os matemáticos precisaram arranjar um modo mais rápido de transmitir resultados importantes para a sua comunidade, e essa era a função dos pré-impressos. Atualmente, o pré-impresso vai para a internet. São sempre provisórios; a publicação plena requer revisão dos pares. Os pré-impressos ajudam nesse processo, porque qualquer um pode lê-los, procurar erros ou aperfeiçoamentos e comunicar aos autores. Na verdade, a versão publicada muitas vezes difere consideravelmente do pré-impresso, exatamente por esse motivo.

A prova final levou mil horas de tempo de computador e envolveu 487 regras de descarga; os resultados foram publicados em dois artigos com um suplemento de 450 páginas mostrando todas as 1.482 configurações. Naquele tempo, foi um *tour de force*.

A PRINCIPAL REAÇÃO da comunidade matemática mais ampla, contudo, foi de vago desapontamento. Não com o resultado; não com a notável façanha computacional. O que era decepcionante era o método. Na década de 1970, as provas matemáticas eram coisas que você escrevia à mão e conferia à mão. Como eu disse no Capítulo 1, uma prova é uma história cujo enredo nos convence de que a afirmação é verdadeira. Mas essa história não tinha enredo. Ou, se tivesse, havia um grande furo no meio:

Era uma vez uma linda conjectura. Sua mãe lhe dizia para nunca entrar na escura e perigosa floresta. Mas um dia a Pequena Conjectura Quatro Cores saiu de mansinho e foi vagar dentro da floresta inevitável. Sabia que se toda configuração na floresta fosse redutível, ela teria uma prova, se transformaria na Pequena Teorema Quatro Cores e seria publicada numa revista dirigida pelo Príncipe Ton. Ela se deparou com um computador coberto de doce, bem no meio da floresta, e dentro havia um Lobo disfarçado de programador. E o Lobo disse "Sim, são todas redutíveis", e todos viveram felizes para sempre.

Não, não dá certo. Estou sendo petulante, mas o furo desse conto de fadas é igual ao furo na prova Appel-Haken, ou, pelo menos, o que a maioria dos matemáticos considera como sendo o furo na prova. *Como sabemos que o Lobo está certo?*

Rodamos o nosso próprio computador e descobrimos se está de acordo. Mas por mais vezes que façamos isso, não tem o mesmo toque de autenticidade que, digamos, a minha prova de que não se pode cobrir um tabuleiro de xadrez mutilado com dominós. Não se pode capturá-la por inteiro. Não seria possível verificar todos os cálculos à mão mesmo se você vivesse 1 bilhão de anos. Pior, você não acreditaria na resposta se pudesse. Os seres humanos cometem erros. Em 1 bilhão de anos, cometem um monte de erros.

Computadores, de forma geral, não cometem. Se um computador e uma pessoa fizerem, ambos, uma operação aritmética realmente complicada e

os resultados não baterem, a aposta boa seria no computador. Mas não é certeza. Um computador que esteja funcionando exatamente como projetado pode cometer um erro; por exemplo, um raio cósmico pode cruzar sua memória e mudar um 0 para 1. Você pode se proteger disso fazendo novamente a computação, porém, mais seriamente, os projetistas podem cometer erros. O chip Intel 5 Pentium tinha um erro em suas rotinas para aritmética de ponto flutuante: caso se pedisse para dividir 4195835 por 3145727, ele respondia com 1,33373, quando a resposta correta é 1,33382. Aparentemente, quatro entradas numa tabela haviam sido deixadas de fora.[7] Outras coisas que podem dar errado incluem o sistema operacional do computador e bugs no programa do usuário.

Um bocado de ar quente filosófico foi gasto na proposição de que a prova de Appel-Haken usando o computador mudava a natureza da "prova". Posso ver aonde os filósofos querem chegar, mas o conceito de prova que os matemáticos que trabalham utilizam não é o mesmo que ensinamos aos alunos de graduação em aulas de lógica matemática. E mesmo quando o conceito mais formal se aplica, nada exige que a lógica de cada passo seja verificada por um ser humano. Durante séculos, os matemáticos usaram máquinas para aritmética rotineira. E mesmo que uma pessoa repasse realmente uma prova linha por linha, sem encontrar erros, como sabemos que ela não deixou passar algum? Lógica perfeita, incontestável, é um ideal que almejamos. Humanos imperfeitos fazem o melhor que podem, mas nunca podem remover todo elemento de incerteza.

Em *Four Collours Suffice* (Quatro cores bastam), Robin Wilson põe o dedo num aspecto sociológico básico da reação da comunidade:

A plateia dividia-se em dois grupos: os com mais de quarenta anos não podiam ser convencidos de que uma prova por computador estava correta, enquanto os com menos de quarenta não podiam ser convencidos de que uma prova contendo setecentas páginas de cálculos manuais pudesse estar correta.

Se nossas máquinas são melhores do que nós em algumas coisas, faz sentido usar máquinas. *Técnicas* de prova podem mudar, mas de qualquer maneira isso ocorre o tempo todo: é chamado de "pesquisa". O conceito de prova não fica radicalmente alterado se alguns passos forem dados por um computador. Uma prova é uma história; uma prova assistida por computador é uma história comprida demais para ser contada por inteiro,

então é preciso contentar-se com uma sinopse e um enorme apêndice automatizado.

DESDE O TRABALHO pioneiro de Appel e Haken, os matemáticos acostumaram-se à assistência do computador. Eles ainda *preferem* provas que se baseiem unicamente na capacidade cerebral humana, mas a maioria deles não faz mais essa exigência. Nos anos 1990, porém, ainda havia certa dose de desconforto justificável em relação à prova de Appel-Haken. Assim, em vez de reconferir o trabalho, alguns matemáticos decidiram refazer a prova toda, tirando proveito de novos avanços teóricos e computadores muito mais modernos. Em 1994, Neil Robertson, Daniel Sanders, Paul Seymour e Robin Thomas jogaram fora tudo o que continha o artigo de Appel-Haken, exceto a estratégia básica. Em um ano tinham descoberto um conjunto inevitável de 633 configurações, cada uma delas podendo ser provada como redutível usando apenas 32 regras de descarga. Isso era muito mais simples do que as 1.482 configurações e 487 regras de descarga de Appel e Haken. Os computadores de hoje são tão rápidos que agora a prova inteira pode ser verificada em um computador doméstico em poucas horas.

Está tudo muito bem, mas o computador ainda é rei. Podemos nos livrar dele? Há uma crescente sensação de que, neste caso particular, uma história que pode ser entendida completamente pelos homens não pode ser totalmente inconcebível. Talvez novas percepções do problema das quatro cores acabem levando a uma prova mais simples, com pouca ou nenhuma assistência do computador, de modo que os matemáticos possam lê-la, pensar a seu respeito e dizer "Sim!". Ainda não conhecemos essa prova, e pode ser que não exista, mas uma sensação no ar…

Os matemáticos estão aprendendo um bocado sobre grafos. Topólogos e geômetras estão encontrando relações profundas entre grafos e áreas da matemática absolutamente diferentes, inclusive algumas que se aplicam à física matemática. Um dos conceitos que reaparecem, de tempos em tempos, é a curvatura. O nome é adequado: a curvatura do espaço diz quanto ele é curvo. Se é achatado como um plano, a curvatura é zero. Se ele se curva numa mesma direção, do mesmo modo que um morro curva-se para baixo em todos os lados, tem curvatura positiva. E se for como uma passagem nas montanhas, curvada para cima em algumas direções e para baixo em outras, sua curvatura é negativa. Há teoremas geométricos, descendentes da fórmula de Euler, que relacionam grafos desenhados em

um espaço com a própria curvatura do espaço. A fórmula de Heawood para um toro com g furos já aponta para isso. Uma esfera tem curvatura positiva, um toro representado como quadrado com bordas opostas correspondentes (Figura 12, direita) tem curvatura zero, e um toro com dois ou mais furos tem curvatura negativa. Logo, existe algum tipo de elo entre curvatura e colorir mapas.

Por trás desse elo está uma característica proveitosa da curvatura: é difícil livrar-se dela. É como um gato debaixo do tapete. Se o tapete está plano, não há gato, mas se você vê uma corcova, há um gato por baixo. Você pode caçar o gato pelo tapete, mas tudo que consegue é mover a corcova de um lado para outro. De maneira similar, a curvatura pode ser movida, mas não removida. A menos que o gato chegue até a borda do tapete, e nesse caso ele pode escapar, levando a curvatura junto. As regras de descarga de Heesch são um pouco como uma curvatura disfarçada. Elas mudam a carga elétrica de um lugar para outro, mas não a destroem. Poderia existir algum conceito de curvatura para um grafo e alguma ardilosa regra de descarga que, com efeito, force a curvatura de um lado para outro?

Se existir, você poderia ser capaz de persuadir um grafo a colorir a si mesmo automaticamente. Atribua uma curvatura a seus pontos (e talvez traços); então deixe o grafo redistribuir a curvatura mais regularmente. Talvez aqui “regularmente” signifique que quatro cores bastam, se fizermos toda a montagem inicial corretamente. É apenas uma ideia, não é minha, e eu não a expliquei em detalhes suficientes para que faça muito sentido. Mas reflete alguma intuição matemática, e oferece esperança de que seja possível encontrar ainda uma prova mais conceitual do teorema das quatro cores – um conto arrebatador em vez de um resumo com 1 bilhão de listas telefônicas como apêndice. Vamos encontrar uma ideia semelhante, num contexto muito mais sofisticado, no Capítulo 10, e ela resolveu um problema ainda maior em topologia.

5. Simetria esférica

A conjectura de Kepler

TUDO COMEÇOU COM um floco de neve.

A neve tem uma beleza estranha. Cai do céu em flocos brancos fofos, é soprada pelo vento para criar montinhos e morros macios que cobrem a paisagem, toma espontaneamente formas estranhas. É fria. Pode-se esquiar nela, andar de trenó, fazer bolas e bonecos de neve… E, se tiver azar, ficar soterrado por toneladas dela. Quando se vai, não volta para o céu – não diretamente como flocos brancos. Transforma-se em simples água comum, que pode evaporar e voltar para o céu, é claro, mas poderá viajar, descendo pelos rios e percorrendo um comprido caminho até o mar, e passar um longo tempo nos oceanos. Neve é uma forma de gelo, e gelo é água congelada.

Isso não é novidade. Deve ter sido óbvio para os Neandertais.

Flocos de neve não são absolutamente bolotas disformes. Quando estão prístinos, antes de começarem a derreter, muitos deles são minúsculas e intrincadas estrelas: planos, seis lados e simétricos. Outros são simples hexágonos. Alguns têm menos simetria, outros uma substancial terceira dimensão, mas flocos de neve hexagonais são icônicos e difundidos. Isso tampouco é novidade: basta que você ao vê-los reconheça um cristal. Mas estes não são cristais comuns, com faces planas, poligonais. Sua característica mais intrigante adiciona uma pitada de caos: apesar de ter a mesma simetria, a estrutura detalhada difere de um floco de neve para outro. Não existem dois flocos de neve iguais, é o que dizem. Sempre me perguntei como podem sabê-lo, mas os números favorecem esse ponto de vista, se você é pedante o suficiente em relação ao que conta para ser igual.

Por que os flocos de neve têm seis lados? Quatrocentos anos atrás, um dos grandes matemáticos e astrônomos do século XVII fez-se esta pergunta, e pensou seu caminho para uma resposta, que acabou sendo surpreendente, ainda mais pelo fato de ele não ter realizado nenhum experimento especial. Apenas juntou algumas ideias simples que eram conhecidas de todo mundo. Da mesma maneira que sementes de romã estão empacotadas dentro da fruta.

Seu nome era Johannes Kepler, e ele tinha uma boa razão para pensar sobre flocos de neve. Seu sustento dependia de um rico patrocinador, John Wacker de Wackenfels. Naquela época, Kepler era o matemático da corte do sacro imperador romano Rodolfo II, e Wacker, um diplomata, era conselheiro do soberano. Kepler queria dar a seu patrocinador um presente de ano-novo. O ideal era que fosse barato, inusitado e estimulante. Deveria dar a Wacker uma percepção das notáveis descobertas que seu dinheiro estava possibilitando. Assim, Kepler coletou suas ideias sobre flocos de neve num pequeno livro, e esse foi o presente. O título era *De Nive Sexangula* (Dos flocos de neve sextavados). A data era 1611. Escondido dentro do livrinho, um dos principais passos no pensamento de Kepler, estava um breve comentário: uma charada matemática que não seria resolvida por 387 anos.

KEPLER ERA UM INVETERADO investigador de padrões. Seu trabalho científico mais influente foi a descoberta das três leis básicas do movimento planetário, sendo a primeira e mais conhecida a de que a órbita é uma elipse. Era também um místico, profundamente imerso na concepção de mundo pitagórica de que o universo se baseia em números, padrões e formas matemáticas. Dedicava-se à astrologia, bem como à astronomia: os matemáticos frequentemente faziam serão como astrólogos naquela época, porque na realidade podiam somar para descobrir quando Aquário estava em ascendente. Patrocinadores abastados, mesmo a realeza, lhes pagavam para fazer horóscopos.

Em seu livro, Kepler ressaltava que a neve começa como vapor d'água, que não tem forma, e de alguma maneira transforma-se em flocos sólidos de seis lados. Algum agente deve causar essa transição, insistia Kepler:

Teria [este agente] imprimido a forma sextavada na matéria como a matéria exigia, ou a partir de sua própria natureza – uma natureza, por exemplo, na qual esteja inata ou a ideia da beleza inerente ao hexágono ou o conhecimento do propósito que essa forma subserve?

Na busca da resposta, ele considerou outros exemplos de formas hexagonais na natureza. Favos de mel em colmeias de abelhas lhe vieram à mente. Elas são feitas de duas camadas de células hexagonais, lado a lado, e suas extremidades comuns são formadas por três losangos –

paralelogramos com os quatro lados iguais. Essa forma lembrou Kepler de um sólido chamado rombidodecaedro (Figura 15). Ele não é um dos cinco sólidos regulares que os pitagóricos conheciam e que Euclides classificou, mas tem uma propriedade especial: cópias idênticas podem se encaixar perfeitamente de modo a preencher o espaço sem vazios. A mesma forma ocorre em romãs, onde pequenas sementes redondas crescem espremendo-se umas contra as outras, e são portanto forçadas a criar um empilhamento eficiente.

Como qualquer matemático sensato, Kepler começa com o caso mais simples, no qual esferas formam uma camada plana única. Isso equivale a arrumar círculos idênticos no plano. Aqui ele encontra apenas dois arranjos regulares. Em um deles, as esferas estão arranjadas em quadrados (Figura 16, esquerda); no outro, em triângulos equiláteros (Figura 16, direita). Esses arranjos, repetidos ao longo de todo o plano infinito, são o reticulado quadrado e o reticulado triangular. A palavra “reticulado” refere-se ao seu padrão espacialmente periódico, que se repete em duas direções independentes. As figuras necessariamente só mostram uma porção finita do padrão, de modo que as bordas devem ser ignoradas. O mesmo vale para as Figuras 17 a 20. A Figura 16 (esquerda e direita) mostra cinco fileiras de esferas, e em cada fileira elas tocam suas vizinhas. No entanto, o reticulado triangular é ligeiramente espremido: suas fileiras estão mais juntas. Assim, as esferas no reticulado triangular estão organizadas mais densamente do que as do reticulado quadrado.

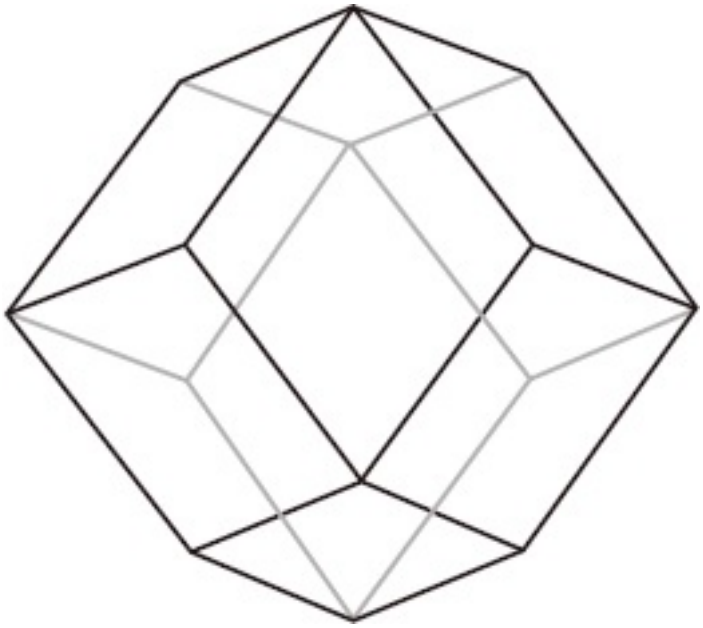


FIGURA 15 O rombidodecaedro, um sólido com doze faces de losangos.

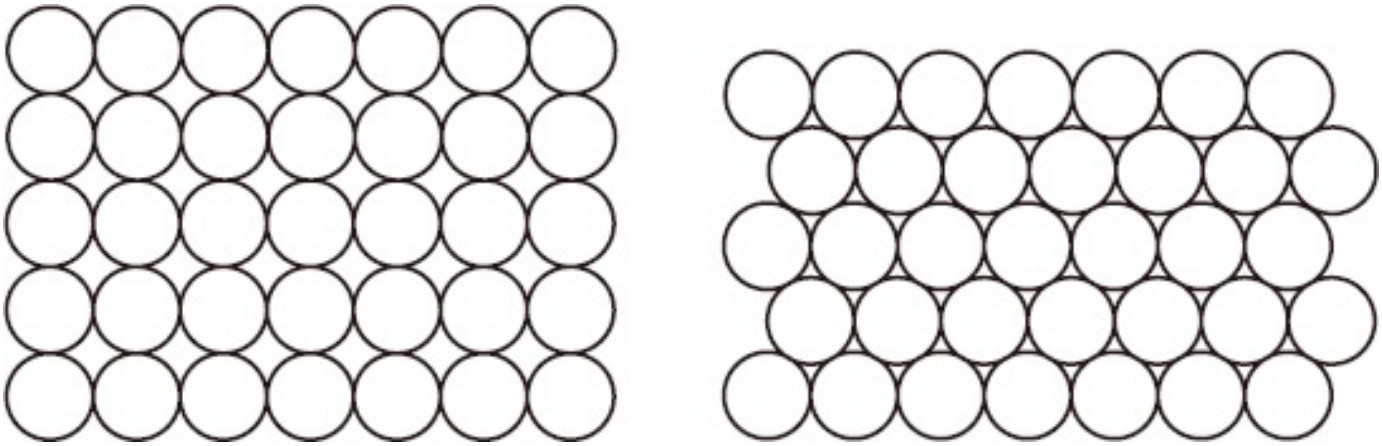


FIGURA 16 *Esquerda*: Empilhamento em reticulado quadrado. *Direita*: Empilhamento em reticulado triangular (também chamado hexagonal).

A seguir Kepler pergunta como camadas sucessivas desse tipo podem ser colocadas umas sobre as outras, e considera quatro casos. Para os dois primeiros, todas as camadas são reticulados quadrados. Um modo de empilhar as camadas é colocar as esferas em cada camada diretamente sobre as da camada abaixo. Então, cada esfera terá seis vizinhas imediatas: quatro na sua camada, uma acima e uma abaixo. Esse empilhamento é como um tabuleiro de xadrez tridimensional formado de cubos, e é nisso que se transformaria se fosse possível expandir as esferas até não poderem expandir-se mais. Mas isso, diz Kepler, "não será o empilhamento mais apertado". Ela pode ficar mais apertada deslizando a segunda camada para o lado, de modo que as esferas se encaixem perfeitamente nas reentrâncias entre as esferas da camada inferior (Figura 17, esquerda). Repita o processo, camada por camada (Figura 17, direita). Agora cada esfera tem doze vizinhas: quatro na sua própria camada, quatro acima e quatro abaixo. Se você inflá-las ao máximo preencherá os espaços com rombidodecaedros.

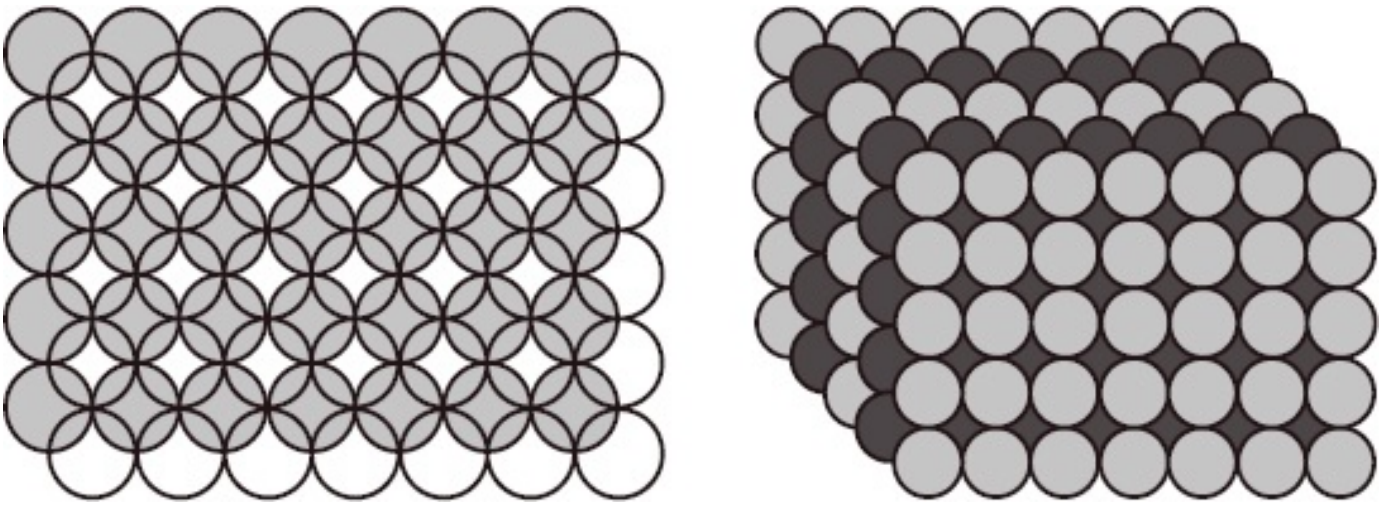


FIGURA 17 *Esquerda*: Acrescenta-se uma segunda camada de esferas (círculos vazados) em cima da primeira camada (cinza). *Direita*: Repete-se esta construção.

Nos dois outros casos, as camadas são reticulados triangulares. Se forem empilhadas de maneira que as esferas em cada camada estejam diretamente acima das que se encontram abaixo, então cada esfera terá oito vizinhas: seis na sua camada, uma acima e uma abaixo. Como alternativa, as esferas na camada seguinte podem ser novamente encaixadas nos vãos entre as esferas da camada inferior. Agora cada esfera tem doze vizinhas: seis na sua própria camada, três acima e três abaixo. Esse é o mesmo número de vizinhas que as das esferas do segundo arranjo de reticulados quadrados, e Kepler fornece uma análise cuidadosa da geometria para mostrar que esse quarto arranjo é, na verdade, o mesmo que o segundo. A única diferença é que as camadas quadradas não são mais horizontais, mas têm um ângulo de inclinação. Ele escreve: "Logo, no empilhamento mais apertado em três dimensões, o padrão triangular, não pode existir sem o quadrado, e vice-versa." Voltarei a isso: é importante.

Tendo determinado a geometria básica do empilhamento das esferas, Kepler retorna ao floco de neve e sua simetria hexagonal. Ele se recorda do empilhamento de esferas em reticulado triangular num plano, no qual cada esfera é cercada por seis outras, formando um hexágono perfeito. Esse, conclui ele, deve ser o motivo pelo qual os flocos de neve são sextavados.

Este capítulo não trata fundamentalmente de flocos de neve, mas a explicação de Kepler a respeito da sua simetria é muito semelhante àquela que daríamos hoje, então seria uma pena parar por aqui. Por que são – como *podem* ser – tão diversos e todavia simétricos? Quando a água se cristaliza para formar gelo, os átomos de hidrogênio e oxigênio que formam as moléculas de água agrupam-se numa estrutura simétrica, o reticulado cristalino. Este reticulado é mais complicado que qualquer arranjo de esferas de Kepler, mas sua simetria dominante é hexagonal. Um floco de neve cresce de uma minúscula "semente" com apenas alguns poucos átomos, arranjados como um pequeno pedaço de reticulado. Essa semente tem a mesma simetria hexagonal e prepara a cena para o crescimento do cristal de gelo quando o vento, dessa maneira, a sopra para dentro de uma nuvem de tempestade.

A grande variedade de padrões de flocos de neve é uma consequência das diversas condições dentro da nuvem. Dependendo da temperatura e da umidade, o crescimento do cristal pode ser uniforme, com átomos sendo

adicionados na mesma proporção ao longo de toda a fronteira, levando a hexágonos de lados retos, ou pode ser dentrítico, com uma taxa de crescimento que varia de um lugar para outro, resultando em estruturas semelhantes a árvores. Quando o floco em crescimento é transportado para lá e para cá através da nuvem, essas condições continuam mudando, aleatoriamente. Mas o floco é tão pequeno que em qualquer dado momento as condições são essencialmente as mesmas em todos os seis cantos. Logo, todos fazem a mesma coisa. Todo floco de neve carrega vestígios de sua história. Na prática, a simetria hexagonal nunca é exata, mas geralmente é bem próxima. O gelo é um material estranho, e outras formas também são possíveis – cravos, placas planas, prismas hexagonais, prismas com placas nas pontas. A história completa é muito complicada, mas tudo se articula em torno de como os átomos estão arranjados em um cristal de gelo.[1] No tempo de Kepler, a teoria atômica era, na melhor das hipóteses, uma sugestão vaga de alguns gregos antigos; é impressionante como ele foi longe com base em observações folclóricas, experimentos mentais e um sentido de padrão.

A CONJECTURA DE KEPLER não trata de flocos de neve como tais. É a sua observação espontânea de que empilhar camadas de esferas em arranjos apertados, de modo que camadas sucessivas se encaixem nos vãos entre as esferas da camada abaixo, leva ao "empilhamento mais apertado em três dimensões". A conjectura pode ser resumida informalmente: se você quer colocar um grande número de laranjas numa caixa grande, preenchendo o máximo possível da caixa, então você deve arrumá-las do jeito que qualquer merceeiro o faria.

A dificuldade não é encontrar a *resposta*. Kepler nos diz qual é a resposta. O difícil é provar que ele estava certo. No decorrer dos séculos, acumulou-se uma profusão de evidências indiretas. Ninguém foi capaz de propor um empilhamento mais próximo. O mesmo arranjo de átomos é comum em cristais, onde um empilhamento eficiente presumivelmente corresponde à mínima energia, um princípio padrão que governa muitas formas naturais. Esse tipo de evidência bastava para satisfazer a maioria dos físicos. Por outro lado, ninguém conseguia apresentar uma prova de que *não havia* nada melhor. Questões simples do mesmo tipo, como arrumar círculos num plano, acabavam revelando profundezas ocultas. Toda a área era difícil e cheia de surpresas. Tudo isso preocupava os matemáticos, mesmo que a maioria deles pensasse que Kepler dera a

resposta certa. Em 1958, C. Ambrose Rogers descreveu a conjectura de Kepler como algo que “muitos matemáticos acreditam e todos os físicos sabem”.[2] Este capítulo descreve como os matemáticos transformaram a crença em certeza.

Para entender o que fizeram, precisamos dar uma boa olhada no arranjo de esferas de Kepler, que é conhecido como reticulado cúbico de faces centradas. Quando fazemos isso, as sutilezas do problema começam a aparecer. A primeira pergunta que nos vem à cabeça é por que usamos camadas quadradas. Afinal, o empilhamento mais ajustado numa camada só ocorre no reticulado *triangular*. A resposta é que também podemos obter o reticulado cúbico de faces centradas usando camadas triangulares; essa é a essência do comentário de Kepler de que “o padrão triangular não pode existir sem o quadrado”. No entanto, é mais fácil descrever o reticulado cúbico de faces centradas usando camadas quadradas. Como bônus, vemos que a conjectura de Kepler não é tão direta e óbvia quanto um merceeiro arrumando laranjas num caixote.

Suponha que comecemos com uma camada plana de esferas arranjadas em triângulos (Figura 16, direita). Entre as esferas existem reentrâncias curvas triangulares, e uma outra camada de esferas pode se encaixar nessas reentrâncias. Quando começamos com uma camada quadrada, fomos capazes de usar todas as reentrâncias, de modo que a posição da segunda camada, e a das seguintes, era determinada de modo único. Esse não será mais o caso se começarmos com um arranjo triangular. Não podemos usar todas as reentrâncias, porque estão próximas demais. Podemos usar apenas metade delas. Uma escolha é mostrada na Figura 18 (esquerda), usando todos os pequenos pontos cinzentos, e a Figura 18 (direita) mostra como deve ser colocada a camada seguinte de esferas. A segunda maneira de encaixar uma camada nova nas reentrâncias da camada 1 é mostrada na Figura 19 (esquerda) usando pontos mais escuros. Esses pontos coincidem com as reentrâncias da camada 2, então adicionamos uma camada 3 nas posições correspondentes: o resultado é a Figura 19 (direita).

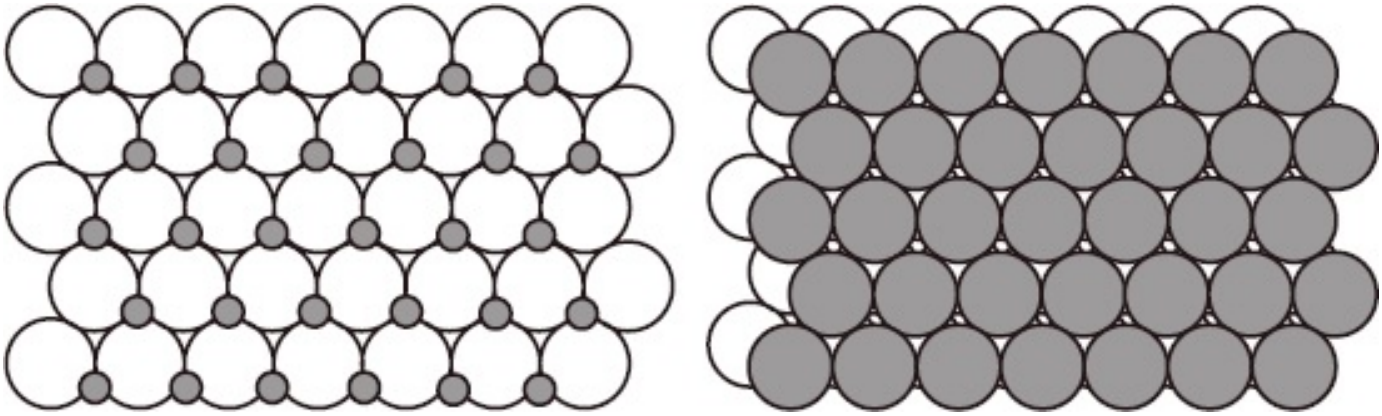


FIGURA 18 Encaixe de um reticulado triangular num conjunto de vãos da camada inferior.

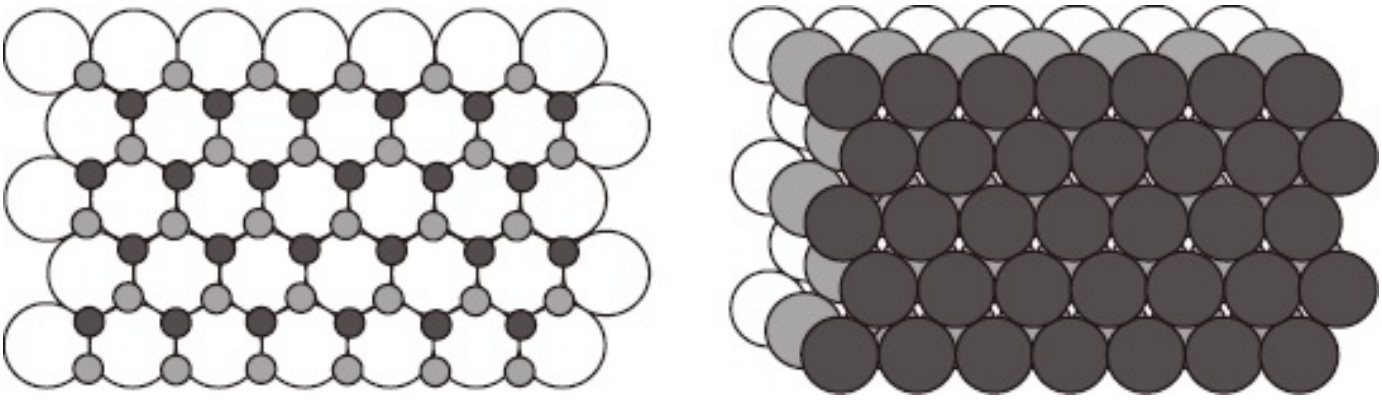


FIGURA 19 Empilhando reticulados triangulares um sobre o outro.

A distinção entre essas escolhas na verdade não faz diferença alguma quando se tem apenas duas camadas. Se girarmos o segundo arranjo em sessenta graus, obteremos o primeiro. Eles são a mesma coisa “até em termos de simetria”. Mas após as primeiras duas camadas terem sido posicionadas, há duas escolhas genuinamente distintas para a terceira camada. Cada nova camada tem dois sistemas de vãos, mostrados pelos pontos claros e escuros na Figura 19 (esquerda). Um sistema se sobrepõe exatamente aos centros da camada imediatamente abaixo, visíveis como pequenos triângulos cinza-claro na Figura 19 (direita). O outro sobrepõe-se aos vãos na camada abaixo dela, visíveis como triângulos contendo um minúsculo hexágono branco na Figura 19 (direita). Para obter o reticulado cúbico de faces centradas devemos usar as posições escuras para a terceira camada, e então continuar com o mesmo padrão indefinidamente.

Não é inteiramente óbvio que o resultado seja o reticulado cúbico de faces centradas. Onde estão os quadrados? A resposta é que estão presentes, mas inclinados num certo ângulo. A Figura 20 mostra seis camadas triangulares sucessivas, com certa quantidade de esferas

removidas. As setas indicam as linhas e colunas do reticulado quadrado, escondido no interior. Camadas paralelas a este também são reticulados quadrados, e encaixam-se exatamente na forma que usei para construir o reticulado cúbico de faces centradas.

Quão "apertado" é esse empilhamento? Nós medimos o grau de aperto (eficiência, proximidade) de um empilhamento pela sua densidade: a proporção de espaço ocupado pelas esferas.[3] Quanto maior a densidade, mais apertado, espremido é o empilhamento. Cubos juntam-se com densidade 1, ocupando todo o espaço. Esferas obviamente precisam deixar vazios, de modo que a densidade é menor que 1. Para o reticulado cúbico de faces centradas, a densidade é exatamente $\pi/\sqrt{18}$, aproximadamente 0,7405. Logo, para este empilhamento, as esferas preenchem pouco menos que três quartos do espaço. A conjectura de Kepler afirma que nenhum empilhamento de esferas pode ter densidade maior que esta.

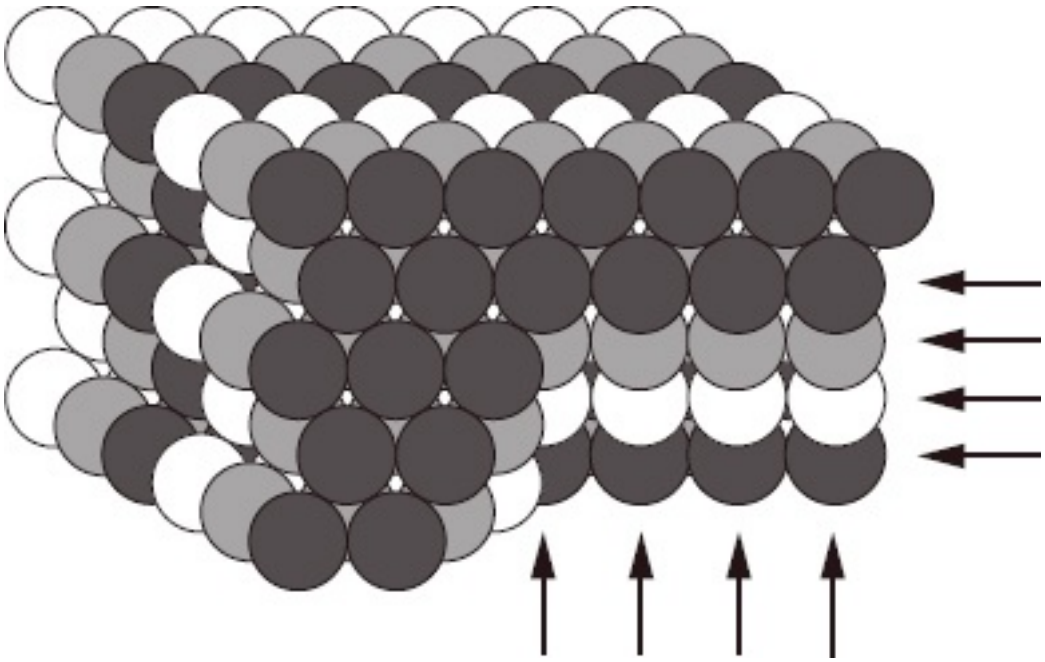


FIGURA 20 Escondidas dentro das camadas triangulares estão camadas quadradas, em posição oblíqua.

Eu fiz essa afirmação com muito cuidado. Não disse "o empilhamento cúbico com face no centro tem densidade maior que qualquer outro". Isso é falso, espetacularmente falso. Para ver por quê, volte à construção do reticulado cúbico de faces centradas usando camadas triangulares. Eu disse que, uma vez determinadas as duas primeiras camadas, há duas escolhas para a terceira camada: o reticulado cúbico de faces centradas aparece se usarmos a segunda opção, os pontos escuros. O que acontece se usarmos a

outra, os pontos cinza-claro? Agora a camada 3 assenta-se exatamente em cima da camada 1. Se continuarmos neste padrão, colocando cada nova camada exatamente acima da camada que está dois níveis abaixo, obteremos um segundo reticulado de empilhamento: o reticulado hexagonal, que é genuinamente diferente do empilhamento cúbico com face centrada, mas tem a mesma densidade. Isso é intuitivamente claro porque as duas maneiras de se colocar a terceira camada estão relacionadas por uma simetria de rotação, então ela se encaixa na camada anterior com o mesmo grau de adensamento, de um modo ou de outro.

Essas são as duas únicas arrumações *de reticulado* que podem ser obtidas a partir de camadas triangulares sucessivas, mas em 1883 o geólogo e cristalógrafo William Barlow mostrou que podemos escolher a localização de cada camada sucessiva ao acaso a partir das duas possibilidades. Como cada posição faz a mesma contribuição para a densidade, todas essas arrumações têm densidade $\pi/\sqrt{18}$. Há infinitas sequências aleatórias, levando a infinitas arrumações diferentes, todas com a mesma densidade.

EM SUMA, não existe algo como "o" empilhamento de esferas mais denso. Em vez disso, elas são infinitas, todas igualmente densas. Essa ausência de exclusividade é um aviso: este não é um problema direto e imediato. Se Kepler estava certo, a *densidade* ótima é única, mas infinitos arranjos têm essa densidade. Assim, a prova de que essa é realmente a densidade ótima não é apenas uma questão de encaixar camadas sucessivas de esferas da forma mais apertada possível. Existem opções.

Por mais impressionante que seja a experiência dos donos de mercearias – e o reticulado cúbico de faces centradas seguramente estava presente nos mercados egípcios pré-dinásticos – nem de longe ela é conclusiva. Na verdade, é meio por acaso que o método da mercearia traga uma boa resposta. O problema com que os merceeiros se deparam não é o de arrumar laranjas da forma mais apertada possível no espaço, onde qualquer arranjo é em princípio possível. É empilhar as laranjas de maneira estável, num mundo onde o chão é plano e a gravidade age para baixo. Os merceeiros começam naturalmente por fazer uma camada; aí acrescentam outra; e assim por diante. Se estiverem pondo as laranjas num caixote retangular, é provável que formem um reticulado quadrado. Se as laranjas não estiverem confinadas, então um reticulado quadrado ou triangular é

natural. Acontece que ambas dão o mesmo reticulado cúbico de faces centradas – pelo menos, se as camadas forem colocadas apropriadamente no caso triangular. Na realidade, o reticulado quadrado dá a impressão de uma má escolha, por não parecer a maneira mais densa de arrumar as camadas. Por sorte ou julgamento, isso acaba não tendo importância.

Físicos não estão interessados em laranjas. O que eles querem arrumar são átomos. Um cristal é um arranjo regular, espacialmente periódico, de átomos. A conjectura de Kepler explica a periodicidade como uma consequência natural dos átomos arrumando-se da maneira mais apertada possível. No que diz respeito à maioria dos físicos, a existência de cristais é evidência suficiente, logo a conjectura é claramente verdadeira. No entanto, acabamos de ver que existem infinitas maneiras de se arrumar esferas com densidade igual à do reticulado cúbico de faces centradas e do reticulado hexagonal, nenhuma das quais é espacialmente periódica. Então por que a natureza usa padrões periódicos para os cristais? Uma possível resposta é que não devemos modelar os átomos como esferas.

Matemáticos tampouco estão interessados em laranjas. Como Kepler, preferem trabalhar com esferas perfeitas, idênticas. Eles não consideram convincente o argumento dos físicos. Se não devemos modelar os átomos como esferas, a existência de cristais deixa de ser evidência em favor da conjectura de Kepler. Não dá para ter as duas coisas. Mesmo que você argumente que a conjectura meio que explica o reticulado cristalino, e o reticulado cristalino meio que mostra que a conjectura está correta… existe uma lacuna lógica. Matemáticos querem uma prova.

Kepler não chamou sua afirmação de conjectura: ele simplesmente a colocou no seu livro. Não fica absolutamente claro se ele pretendia que fosse interpretada de maneira tão abrangente. Estaria alegando que o reticulado cúbico de faces centradas era o "empilhamento mais apertado em três dimensões" entre todas as maneiras de arrumar esferas? Ou simplesmente quis dar a entender que era o empilhamento mais apertado entre as três que havia considerado? Não podemos voltar no tempo para perguntar. Qualquer que tenha sido a realidade histórica, a interpretação que interessou aos matemáticos e físicos foi a abrangente, a ambiciosa. Aquela que pede que contemplemos cada maneira possível de se arrumar juntas infinitas esferas num espaço infinito – e mostrar que nenhuma delas tem densidade maior que o reticulado cúbico de faces centradas.

É MUITO FÁCIL SUBESTIMAR a dificuldade da conjectura de Kepler. Seguramente, a maneira de se obter o empilhamento mais apertado é ir acrescentando as esferas uma a uma, fazendo com que cada uma delas encoste no maior número possível de outras esferas à medida que as vamos colocando, não é? Isso inevitavelmente conduz ao padrão de Kepler. E é o que acontece caso você acrescente as esferas na ordem certa, colocando-as nas posições corretas quando houver alternativas. No entanto, não há garantia de que esse processo passo a passo, adicionando as esferas uma de cada vez, não possa ser superado por algo que tenha um alcance muito maior. Qualquer pessoa que tenha arrumado bagagem de férias no porta-malas de um carro sabe que encaixar as coisas uma por uma pode deixar espaços onde não cabe nada, a não ser que se tire tudo e se comece de novo, dessa vez tendo mais cuidado ao se colocar a bagagem. Reconhecidamente, parte do problema de arrumar a bagagem de férias são os diferentes tamanhos e formatos dos objetos, mas o ponto lógico é bastante claro: garantir o empilhamento mais apertado em uma região pequena pode ter efeito contraproducente, impedindo o arranjo mais espremido numa região maior.

Os arranjos que Kepler considerou são muito especiais. É concebível que algum arranjo inteiramente diferente possa arrumar esferas idênticas de maneira ainda mais apertada. Talvez camadas irregulares fossem mais eficientes. Talvez "camadas" seja a ideia errada. E mesmo se estiver absolutamente seguro de que é a ideia correta, você ainda precisa provar.

Não está convencido? Ainda acha que é óbvio? Tão óbvio que *não precisa* de prova? Deixe-me destruir a sua confiança em sua intuição para o empilhamento de esferas. Eis uma questão muito mais simples, envolvendo círculos num plano. Suponha que eu lhe dê 49 círculos idênticos, cada um com diâmetro de uma unidade. Qual é o tamanho do *menor* quadrado capaz de contê-los, se forem arrumados juntos sem sobreposição? A Figura 21 (esquerda) apresenta a resposta óbvia: arrumá-los como garrafas de leite num caixote. O lado do caixote é exatamente sete unidades. Para provar que é a melhor maneira, observe que cada círculo é mantido rigidamente no lugar por todos os outros, de modo que não existe como criar espaço extra. A Figura 21 (direita) mostra que a resposta está errada. Arrume-os da maneira irregular mostrada na figura e eles caberão num caixote cujo lado é ligeiramente menor que 6,98 unidades.[4] Logo, a prova também está errada. Estar rígido não é garantia de que não exista alternativa melhor.

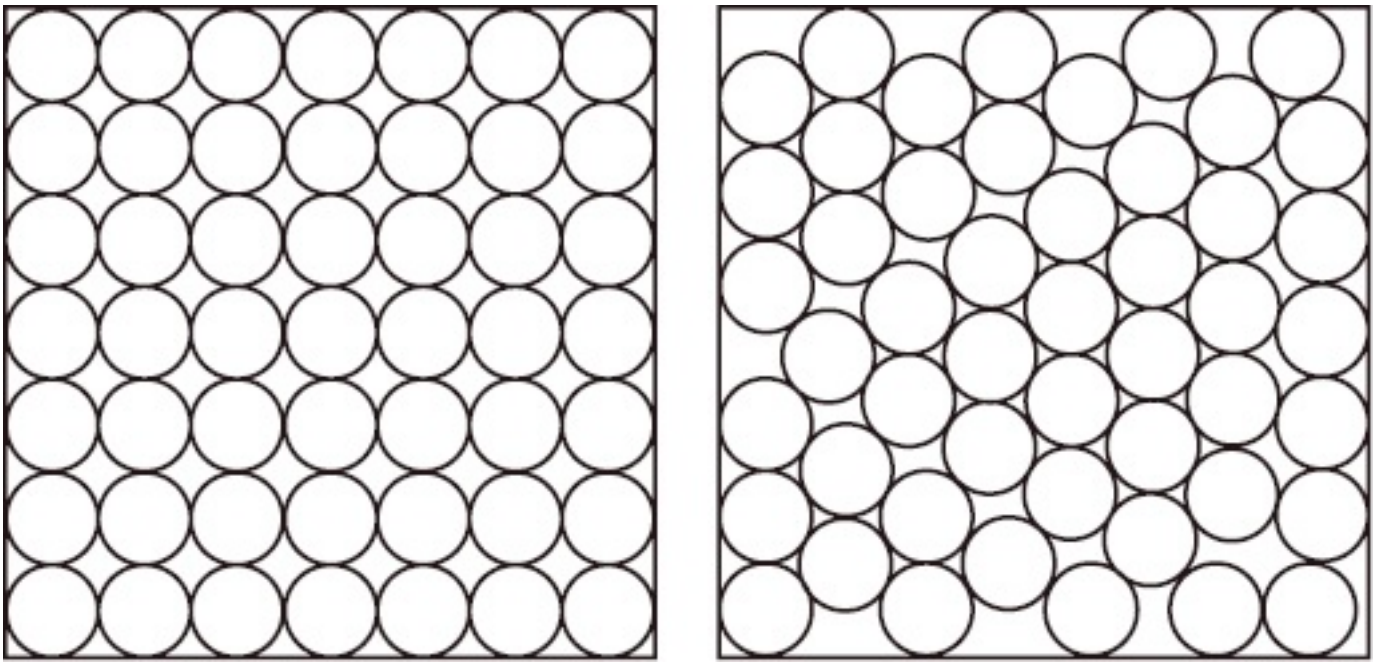


FIGURA 21 *Esquerda*: 49 círculos num quadrado 7 × 7. *Direita*: Como encaixar 49 círculos em um quadrado ligeiramente menor.

Na verdade, é fácil ver que o raciocínio que leva à resposta “7” não pode estar correto. Basta considerar quadrados maiores. Usando um reticulado quadrado, n^2 círculos de diâmetro 1 cabem num quadrado de lado n. Não há meio de melhorar a densidade movendo os círculos de forma contínua, porque o empilhamento é rígido. Mas deve haver arrumações mais densas para um n suficientemente grande, porque um reticulado triangular é mais eficiente que um reticulado quadrado. Se pegarmos um quadrado realmente grande, e encaixarmos o máximo de círculos que conseguirmos usando um reticulado triangular, a vantagem que este tem sobre o reticulado quadrado acabará fazendo com que ele vença, apesar dos “efeitos de borda” nas fronteiras onde temos de deixar vazios. O tamanho da fronteira é $4n$, que fica tão pequeno quanto possível em comparação com n^2. E o ponto exato em que o reticulado triangular assume a liderança é quando $n = 7$. Isso não é óbvio e requer muito trabalho detalhado para se estabelecer, mas algum n precisa funcionar. Rigidez não é suficiente.

Na realidade há duas versões da conjectura de Kepler. Uma considera apenas arrumações em forma de reticulado, onde os centros das esferas formam um padrão especialmente periódico, repetindo-se indefinidamente em três direções independentes como uma espécie de papel de parede sólido. Mesmo assim é um problema difícil, porque há muitos reticulados diferentes no espaço. Os cristalógrafos reconhecem catorze tipos, classificados pelas suas simetrias, e alguns deles são determinados por

números que podem ser ajustados para infinitos valores diferentes. Mas as dificuldades se multiplicam quando consideramos a segunda versão do problema, que permite todas as arrumações possíveis. Cada esfera paira no espaço, não há gravidade, e nenhuma obrigação de formar camadas ou outros arranjos simétricos.

Quando um problema parece difícil demais, os matemáticos o põem de lado e procuram versões mais simples. As ideias de Kepler sobre camadas planas de esferas sugerem começar com empilhamento de círculos no plano. Ou seja, dado um suprimento infinito de círculos idênticos, arrumá-los da maneira mais apertada possível. Agora a densidade é a proporção da *área* que os círculos cobrem. Em 1773, Joseph Louis Lagrange provou que o mais denso empilhamento de círculos em reticulado num plano é o reticulado triangular, com densidade $\pi/\sqrt{12}$ = 0,9069. Em 1831, Gauss estava revisando um livro de Ludwig Seeber, que havia generalizado alguns dos resultados numérico-teóricos de Gauss para equações com três variáveis, e notou que os resultados de Seeber provavam que os reticulados cúbicos com face no centro e hexagonal fornecem o empilhamento de reticulado mais denso no espaço tridimensional. Agora se sabe muitíssima coisa a respeito de arrumações em reticulado em espaços de dimensão superior – 4, 5, 6, e assim por diante. O caso de 24 dimensões é especialmente bem compreendido. (Assim é o tema.) Apesar de parecer impraticável, essa área na verdade tem implicações para a teoria da informação e códigos de computadores.

Arrumações sem ser em reticulado são um assunto completamente diferente. Existem infinitas arrumações desse tipo, e elas não têm nenhuma estrutura regular atraente. Então, por que não ir para o outro extremo e tentar arrumações aleatórias? Em seu livro *Vegetable Staticks*, de 1727, Stephen Hales relatou experimentos nos quais "comprimiu diversos pacotes de ervilhas frescas no mesmo pote", descobrindo que quando estavam todas comprimidas formavam "belos dodecaedros regulares". Ele pareceu ter achado os dodecaedros belos, mas não muito regulares, porque dodecaedros regulares não preenchem o espaço. O que ele viu provavelmente foram rombidodecaedros, que têm sido associados ao empilhamento cúbico com face centrada. G. David Scott colocou montes de rolamentos esféricos num recipiente e o sacudiu, observando que a maior densidade era 0,6366. Em 2008, Chaoming Song, Ping Wang e Hernán Makse deduziram este número analiticamente.[5] Contudo, esse resultado não significa que Kepler estivesse certo – no mínimo porque, como foi

afirmado, implicaria que o reticulado cúbico de faces centradas, cuja densidade é 0,74, não pode existir. O modo mais simples de explicar essa discrepância é que seu resultado ignora exceções extremamente raras: o reticulado cúbico de faces centradas, o reticulado hexagonal e todos os arranjos de camadas triangulares escolhidas aleatoriamente são exceções desse tipo. Da mesma maneira, poderia existir algum outro arranjo com densidade ainda maior. Não pode ser um reticulado, mas uma busca aleatória jamais poderá encontrá-la porque sua probabilidade é zero. Logo, o estudo de arranjos aleatórios, ainda que relevante para muitas questões em física, não nos diz muita coisa a respeito da conjectura de Kepler.

O primeiro avanço realmente significativo veio em 1892, quando Axel Thue proferiu uma palestra para o Congresso Escandinavo de Ciência Natural, esboçando uma prova de que nenhum empilhamento de círculos no plano pode ser mais denso que o reticulado triangular. Sua palestra foi publicada, mas os detalhes são vagos demais para reconstituir a prova que ele teve em mente. Ele apresentou uma nova prova em 1910, que pareceu convincente, salvo por alguns pontos técnicos que ele simplesmente presumiu que pudessem ser resolvidos. Em vez de preencher essas lacunas, László Fejes Tóth obteve uma prova completa por outros métodos em 1940. Logo em seguida, Beniamino Segre e Kurt Mahler acharam provas alternativas. Em 2010, Hai-Chau Chang e Lih-Chung Wang colocaram uma prova mais simples na internet.[6]

ENCONTRAR A MAIOR DENSIDADE para um empilhamento de círculos ou esferas, sob condições específicas, recai numa classe geral de questões matemáticas conhecidas como problemas de otimização. Esse tipo de problema busca o valor máximo ou mínimo de alguma função, que é uma regra matemática para calcular uma grandeza que depende de maneira específica de algum número de variáveis. A regra é frequentemente especificada por uma fórmula, mas isso não é essencial. Por exemplo, o problema do caixote de garrafas de leite, com 49 círculos, pode ser formulado da seguinte maneira. As variáveis são as coordenadas dos centros dos 49 círculos; como cada círculo precisa de duas coordenadas, há 98 variáveis. A função é o tamanho do menor quadrado, com seus lados paralelos aos eixos, que contenha um conjunto dado de círculos sem sobreposição. O problema do caixote de leite equivale a achar um valor mínimo que esta função pode adquirir à medida que as variáveis percorrem todas as arrumações.

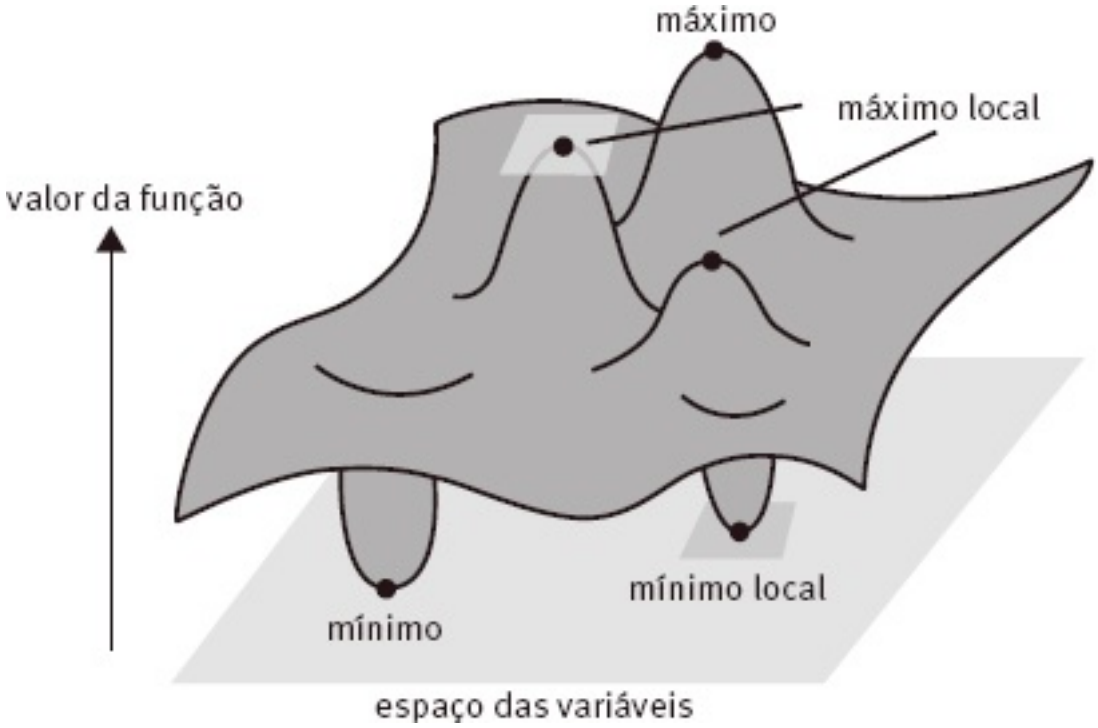


FIGURA 22 Picos e vales de uma função.

Uma função pode ser pensada como uma paisagem multidimensional. Cada ponto da paisagem corresponde a uma escolha das variáveis, e a altitude daquele ponto corresponde ao valor da função. O máximo da função é a altitude do pico mais alto, e o mínimo é a profundidade do vale mais profundo. Em princípio, problemas de otimização podem ser resolvidos por cálculo: a função deve ter uma característica horizontal num pico ou num vale (Figura 22) e o cálculo expressa essa condição como uma equação. Para solucionar o problema do caixote de leite por esse método, temos de resolver um sistema de 98 equações nas 98 variáveis.

Um percalço em problemas de otimização é que equações como essas geralmente têm um grande número de soluções. Uma paisagem possui inúmeros picos locais, mas somente um deles é o mais alto. Pense na cordilheira do Himalaia: mal existe outra coisa *a não ser* picos, ainda assim o Everest detém o recorde de altitude. Métodos para localizar picos, dos quais o mais óbvio é “suba se conseguir”, muitas vezes ficam encalhados num pico local. Outro percalço é que, à medida que o número de variáveis cresce, o mesmo acontece com o provável número de picos locais. Não obstante, esse método às vezes funciona. Mesmo resultados parciais podem ser úteis: se você descobre um pico local, o máximo deve ser no mínimo daquela altitude. É assim que foi encontrado um arranjo melhor no problema do caixote de leite.

Para arrumações em reticulado, a função cujo máximo está sendo buscado depende apenas de uma quantidade finita de variáveis, as direções e comprimentos ao longo dos quais o reticulado se repete. Para arrumações fora do reticulado, a função depende de infinitas variáveis: os centros de todos os círculos ou esferas. Em tais casos o uso direto do cálculo ou outras técnicas de otimização é inútil. A prova de Tóth utiliza uma ideia inteligente para reformular o problema do empilhamento sem reticulado para círculos como problema de otimização em um conjunto *finito* de variáveis. Mais tarde, em 1953, ele percebeu que o mesmo recurso podia ser aplicado em princípio à conjectura de Kepler. Infelizmente, a função resultante depende de cerca de 150 variáveis, um número grande demais para cálculos manuais. Porém, previdentemente, Tóth anteviu uma saída: "Tendo em conta o desenvolvimento rápido dos nossos computadores, é imaginável que o mínimo possa ser determinado com grande exatidão."

Naquela época, a computação ainda estava na primeira infância, e não existia nenhuma máquina suficientemente potente. Assim, o progresso subsequente na conjectura de Kepler seguiu diferentes linhas. Vários matemáticos estabeleceram tetos – limites superiores – a respeito de quanto poderia ser o maior valor da densidade do empilhamento de esferas. Por exemplo, em 1958, Rogers provou que seria no máximo 0,7797: sem exceções raras, este teto aplicava-se a todas as arrumações. Em 1986, J.H. Lindsey melhorou o teto para 0,77844, e em 1988 Douglas Muder aparou um pouquinho mais para obter um teto de 0,77836.[7] Esses resultados mostram que é impossível sair-se *muito* melhor do que o reticulado cúbico de faces centradas, com seu 0,7405. Mas ainda havia um vazio, e pouca perspectiva de livrar-se dele.

EM 1990, Wu-Yi Hsiang, um matemático norte-americano, anunciou uma prova da conjectura de Kepler. Quando os detalhes tornaram-se públicos, porém, rapidamente instalaram-se dúvidas. Quando Tóth resenhou o artigo na *Mathematical Reviews*, escreveu: "Sou indagado [se o artigo fornece] uma prova da conjectura de Kepler, minha resposta é não. Espero que Hsiang preencha os detalhes, mas sinto que a maior parte do trabalho ainda está por fazer."

Thomas Hales, que vinha trabalhando na conjectura há muitos anos, também duvidou que o método de Hsiang pudesse ser reparado. Em vez disso, decidiu que era hora de levar a sério a abordagem de Tóth. Uma nova

geração de matemáticos havia crescido, uma geração para quem recorrer a um computador era mais natural do que recorrer a uma tábua de logaritmos. Em 1996, Hales delineou uma estratégia de prova baseada na ideia de Tóth. Essa estratégia exigia identificar todas as maneiras possíveis de arranjar várias esferas na vizinhança imediata de uma esfera dada. Um empilhamento de esferas é determinado pelos centros das mesmas; para esferas unitárias, estes devem estar separados por ao menos duas unidades. Digamos que duas esferas são *vizinhas* se seus centros distarem, no máximo, 2,51 unidades. Esse valor é uma questão de julgamento: se for pequeno demais, não haverá espaço suficiente para rearranjar vizinhas de modo a melhorar a densidade; se for grande demais, o número de maneiras de arranjar as vizinhas torna-se gigantesco. Hales descobriu que 2,51 era um bom meio-termo. Agora podemos representar como vizinhas estão arranjadas formando uma rede infinita no espaço. Seus pontos são os centros das esferas, e dois pontos são ligados por um traço se forem vizinhos. Essa rede é um tipo de esqueleto do empilhamento, e contém informação vital sobre a vizinhança de cada esfera.

Para qualquer esfera dada, podemos olhar para suas vizinhas na rede e considerar apenas os traços entre essas vizinhas, omitindo a esfera original. O resultado é uma espécie de gaiola cercando o ponto no centro da esfera original. A Figura 23 (par da esquerda) mostra a vizinhança de uma esfera no reticulado cúbico de faces centradas e a gaiola a ela associada. A Figura 23 (par da direita) faz o mesmo para um arranjo especial de esferas, o prisma pentagonal, que acabou se revelando um elemento-chave para a prova. Aqui há duas faixas de pentágonos paralelas ao “equador” da esfera central, mais uma única esfera em cada polo.

As gaiolas formam um sólido com faces planas, e a geometria desse sólido controla a densidade do empilhamento próximo à esfera central.[8] A ideia-chave é associar a cada gaiola um número, conhecido como escore, que pode ser pensado como um meio de avaliar a densidade com que estão empilhadas as vizinhas de uma esfera. O escore não é a densidade em si, mas uma grandeza mais bem-comportada e mais fácil de calcular. Em particular, pode-se achar o escore da gaiola somando-se escores relacionados com suas faces, o que não funciona com a densidade. Em geral, muitas noções diferentes de escore satisfazem essa condição, mas todas concordam numa coisa: para o reticulado cúbico de faces centradas e o reticulado hexagonal o escore é sempre oito “pontos”, não importa que escolha seja feita para sua definição. Aqui um ponto é um número

específico:

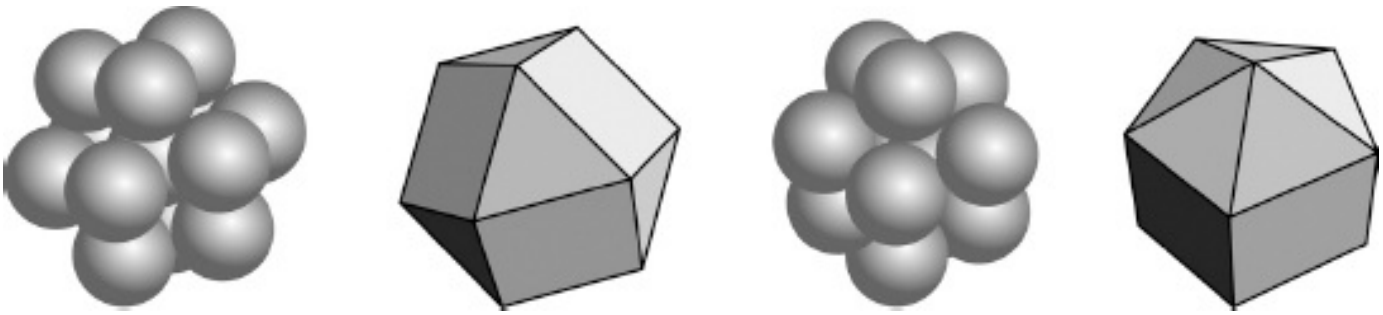


FIGURA 23 *Da esquerda para a direita*: Vizinhança de uma esfera no reticulado cúbico de faces centradas; a gaiola formada pelas suas vizinhas; vizinhança de uma esfera do tipo prisma pentagonal; a gaiola formada pelas suas vizinhas.

$$4 \arctan \frac{\sqrt{2}}{5} - \frac{\pi}{3} = 0{,}0553736$$

Então oito pontos é na realidade 0,4429888. Este curioso número vem da geometria especial do reticulado cúbico de faces centradas. A observação-chave de Hales relaciona a conjectura de Kepler a esse número: se toda gaiola tiver um escore de oito pontos ou menos, a conjectura de Kepler será verdadeira. Então, o foco muda para gaiolas e escores.

Gaiolas podem ser classificadas pela sua topologia: quantas faces elas têm com um dado número de lados, e como essas faces se juntam. Para uma dada topologia, porém, as arestas podem ter muitos comprimentos diferentes. Esses comprimentos afetam o escore, mas a topologia junta uma porção de gaiolas num mesmo grupo, e estas podem ser abordadas da mesma maneira geral. Em sua eventual prova, Hales considerou cerca de 5 mil tipos de gaiola, mas os cálculos principais concentraram-se em algumas centenas. Em 1992, ele propôs um programa de cinco etapas:

1. Provar o resultado desejado quando todas as faces da gaiola são triângulos.
2. Mostrar que os empilhamentos cúbicos de faces centradas e hexagonais têm um escore mais alto que qualquer outra gaiola com a mesma topologia.
3. Lidar com o caso em que todas as faces da gaiola são triângulos e quadriláteros, com exceção do prisma pentagonal, que é mais difícil.
4. Lidar com qualquer gaiola que tenha uma face com mais de quatro lados.
5. Resolver o único caso restante, quando a gaiola é um prisma pentagonal.

A parte 1 foi solucionada em 1994 e a parte 2 em 1995. À medida que o programa evoluiu, Hales modificou a definição de gaiola para simplificar o argumento (seu termo é “estrela de decomposição”). A nova definição não altera as duas gaiolas ilustradas, e não teve quaisquer efeitos sérios sobre as partes da prova que já haviam sido obtidas. Em 1998, usando seu novo conceito, todas as cinco etapas tinham sido completadas. Samuel Ferguson, aluno de Hales, resolveu a parte 5, o caso traiçoeiro do prisma pentagonal.

A análise envolveu intenso uso do computador em todas as etapas. O truque é escolher, para cada rede local, uma noção de escore que torne o cálculo relativamente fácil. Geometricamente, substituir a densidade pelo escore coloca uma espécie de teto por cima da paisagem lisa cujo pico está sendo procurado. O teto é feito de numerosas peças planas (Figura 24). Formas como essa são mais fáceis de trabalhar do que superfícies lisas, porque os máximos devem ocorrer nos vértices, e estes podem ser encontrados resolvendo equações muito mais simples. Existem métodos eficientes para isso, conhecidos como programação linear. Se o teto for construído astuciosamente, de modo que seu pico coincida com o pico da superfície lisa, então essa computação mais simples localiza o pico da superfície lisa.

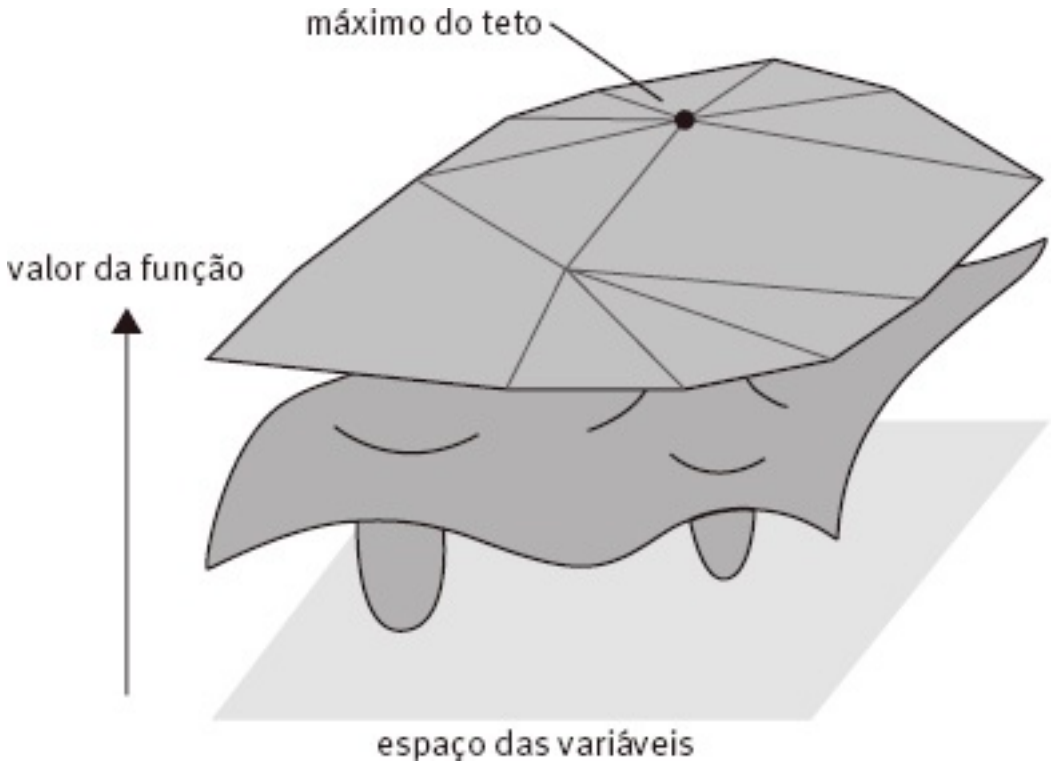


FIGURA 24 Ajustando um teto por cima de uma função.

Há um preço a pagar por essa abordagem: é preciso resolver cerca de

100 mil problemas de programação linear. Os cálculos são longos, mas bem dentro da capacidade dos computadores de hoje. Quando Hales e Ferguson prepararam seu trabalho para publicação, o computador gerou 250 páginas de matemática mais três gigabytes de arquivos de computador.

Em 1999, Hales submeteu a prova à revista *Annals of Mathematics*, e a publicação escolheu um júri de doze especialistas da área. Em 2003, o júri declarou-se "99% certo" de que a prova estava correta. A incerteza restante referia-se aos cálculos do computador; o júri repetira muitos cálculos, além de conferir a maneira como a prova estava organizada e programada, mas foi incapaz de verificar alguns aspectos. Após um adiamento, a revista publicou o artigo. Hales reconheceu que sua abordagem da prova provavelmente jamais seria certificada como 100% correta, então em 2003 anunciou que estava iniciando um projeto para reformular a prova de maneira que pudesse ser verificada por computador usando software de verificação de provas automatizado e padronizado.

Isso pode soar como sair da frigideira para cair no fogo, mas na verdade faz muito sentido. As provas que os matemáticos publicam em revistas especializadas têm a intenção de convencer seres humanos. Como eu disse no Capítulo 1, esse tipo de prova é uma espécie de história. Os computadores são fracos para contar histórias, mas excelentes numa coisa em que somos desastrosos: executar cálculos longos e enfadonhos sem cometer um erro. Eles são ideais para o conceito formal de prova nos livros-textos de graduação: uma série de passos lógicos, cada um consequência dos anteriores.

Cientistas da computação têm explorado essa habilidade. Para conferir uma prova faça o computador verificar cada passo lógico. Deveria ser fácil, mas provas em publicações científicas não são redigidas dessa forma. Elas deixam de fora qualquer coisa óbvia ou rotineira. As frases tradicionais são fáceis de identificar: "É fácil verificar que…" ou "Usando os métodos de Cheesberger e Frits, modificados para levar em conta singularidades isoladas, vemos que…", ou ainda "Um breve cálculo estabelece que…". Computadores (ainda) não conseguem lidar com esse tipo de coisa. Mas os humanos podem reescrever provas preenchendo todas essas lacunas, e aí então os computadores podem verificar cada passo.

O motivo de não estarmos saltando diretamente para dentro do fogo é simples: o software que faz a verificação precisa ser checado apenas *uma*

vez. É um software com propósitos gerais, aplicável a todas as provas e escrito no formato certo. Todas as preocupações sobre provas de computador concentram-se naquela única peça de software. Verifica-se a peça, e ela pode ser usada para verificar todo o resto. Você pode até mesmo "inicializar" o processo escrevendo um software de verificação de provas numa linguagem que possa ser checada por um software de verificação de provas mais simples.

Em anos recentes, provas de muitos teoremas fundamentais da matemática têm sido verificadas dessa maneira. É frequente que as provas tenham de ser apresentadas num estilo mais adequado para a manipulação do computador. Um dos atuais triunfos é uma prova verificada do teorema da curva de Jordan: toda curva fechada no plano que não cruze consigo mesma divide o plano em duas regiões distintas conectadas. Isso pode parecer óbvio, mas os pioneiros da topologia tiveram dificuldades em encontrar uma prova rigorosa. Camille Jordan finalmente conseguiu em 1887 com uma prova de mais de oitenta páginas, mas foi posteriormente criticado por fazer premissas inconsistentes. Em vez disso, o crédito foi para Oswald Veblen, que deu uma prova mais detalhada em 1905, dizendo: "A prova [de Jordan] … é insatisfatória para muitos matemáticos. Ela assume o teorema sem prova no importante caso especial de um polígono simples, e do argumento desse ponto em diante deve-se admitir, no mínimo, que não são dados todos os detalhes." Matemáticos posteriores aceitaram a crítica de Veblen sem objeções, mas recentemente Hales repassou a prova de Jordan e nela "não achou nada de objetável". Na verdade, o comentário de Veblen sobre o polígono é bizarro: o teorema é direto para um polígono, e a prova de Jordan de qualquer maneira não se apoia nessa versão.[9] Provas que contam histórias têm seus próprios perigos. Sempre vale a pena checar se a versão popular da história é a mesma que a original.

Como aquecimento para a conjectura de Kepler, em 2007 Hales deu uma prova formal verificada por computador do teorema da curva de Jordan usando 60 mil linhas de programa de computador. Logo depois, uma equipe de matemáticos produziu outra prova formal usando um software diferente. A verificação por computador não é totalmente infalível, mas as provas tradicionais também não são. De fato, muitos artigos de pesquisa matemática provavelmente contêm um erro técnico em algum ponto. Esses erros revelam-se de tempos em tempos, e a maioria deles acaba mostrando-se inofensiva. Erros sérios geralmente são identificados porque

introduzem inconsistências, de modo que algo visivelmente não faz sentido. Essa é outra desvantagem da narrativa: o preço que pagamos para tornar uma prova compreensível para os humanos é que uma história arrebatadora às vezes pode ser muito convincente mesmo que esteja errada.

Hales chama sua abordagem de *Projeto FlysPecK* – o F, o P e o K significando "*formal proof of Kepler*" ("prova formal de Kepler"). A princípio, ele estimou que levaria cerca de vinte anos para completar a tarefa.[10] Com o projeto em andamento há nove anos, já foi feito um progresso considerável. Pode ser que termine antes.

6. Novas soluções para coisas antigas

A conjectura de Mordell

AGORA ESTAMOS RETORNANDO ao campo da teoria dos números, visando ao último teorema de Fermat. Para preparar o terreno, vou começar com um problema menos familiar, mas indiscutivelmente até mais importante. Em 2002, Andrew Granville e Thomas Tucker o introduziram da seguinte maneira:[1]

Em [1922] Mordell redigiu um dos melhores artigos da história da matemática … Bem no final, Mordell fez cinco perguntas que são essenciais para motivar grande parte da importante pesquisa em aritmética de Diofanto no século XX. A mais importante e mais difícil dessas perguntas foi respondida por Faltings em 1983, ao inventar algumas das mais profundas e poderosas ideias na história da matemática.

Mordell é o britânico Louis Mordell, teórico dos números, nascido nos Estados Unidos em uma família judia de origem lituana, e Faltings é o matemático alemão Gerd Faltings. A referida pergunta veio a ser conhecida como conjectura de Mordell, e a citação já revela seu status atual: provada, brilhantemente, por Faltings.

A conjectura de Mordell pertence a uma importante área da teoria dos números: equações diofantinas. Elas recebem o nome de Diofanto de Alexandria, que escreveu um famoso livro, *Arithmetica*, por volta do ano 250. Acredita-se que originalmente *Arithmetica* incluía treze livros, mas restaram apenas seis, todos em cópias posteriores. Não se tratava de um texto aritmético no sentido de somas e multiplicações. Foi o primeiro texto de álgebra, e reunia a maior parte do que os gregos sabiam a respeito de como resolver equações. Tinha até mesmo uma forma rudimentar de notação algébrica, que se acredita usava a variante ζ da letra grega sigma para a incógnita (nosso x), Δ^{Υ} para seu quadrado (nosso x^2) e K^{Υ} para seu cubo (nosso x^3). A adição era representada colocando-se os símbolos um ao lado do outro, a subtração tinha seu próprio símbolo especial, a inversa da incógnita (nosso $1/x$) era ζ^{χ}, e assim por diante. Os símbolos foram

reconstituídos a partir de cópias e traduções posteriores, que podem não ser totalmente precisas.

No espírito da matemática grega clássica, exigia-se que as soluções de equações buscadas na *Arithmetica* fossem números racionais – isto é, frações como $^{22}/_{7}$ formadas por dois números inteiros. Muitas vezes, exigia-se que *fossem* números inteiros. Todos os números envolvidos eram positivos: os números negativos foram introduzidos vários séculos depois na China e na Índia. Agora chamamos tais problemas de equações diofantinas. O livro inclui alguns resultados notavelmente profundos. De modo particular, Diofanto parece ter tido consciência de que todo número inteiro pode ser expresso como a soma de quatro quadrados perfeitos (inclusive zero). Lagrange deu a primeira prova em 1770. O resultado que nos interessa aqui é a fórmula para todas as trincas pitagóricas, na qual dois quadrados perfeitos somados formam outro quadrado perfeito. O nome vem do teorema de Pitágoras: essa relação vale para os lados de um triângulo retângulo. O exemplo mais conhecido é o celebrado triângulo 3-4-5: $3^2 + 4^2 = 5^2$. Outro é $5^2 + 12^2 = 13^2$. Existe uma infinidade dessas trincas pitagóricas, e há uma receita para encontrá-las em dois lemas (proposições auxiliares) precedendo as Proposições 29 e 30 do Livro X dos *Elementos* de Euclides.

O procedimento de Euclides gera infinitas trincas pitagóricas. Mordell conhecia outras equações diofantinas para as quais existe uma fórmula que produz infinitas soluções. Conhecia também outro tipo de equação diofantina com infinitas soluções, *não* prescritas por uma fórmula. São as chamadas curvas elípticas – um nome bastante tolo, uma vez que elas não têm nada a ver com elipses –, e a infinidade de soluções surge porque quaisquer duas soluções podem ser combinadas para criar outra. O próprio Mordell provou uma das propriedades básicas dessas equações: é preciso apenas um número finito de soluções para gerá-las todas por meio desse processo.

Além desses dois tipos de equação, qualquer outra equação diofantina em que Mordell conseguiu pensar recaía numa dessas duas categorias. Ou era conhecido como tendo apenas um número finito de soluções, inclusive nenhuma, ou ninguém sabia se o número de soluções era finito ou infinito. Isso em si não era novidade, mas Mordell julgou que podia identificar um padrão que ninguém mais havia notado. Não era absolutamente um padrão da teoria dos números; provinha da topologia. O que importava era quantos "furos" a equação tinha. E para isso fazer sentido era preciso

pensar nas soluções em números complexos, não números racionais ou inteiros. O que de certa forma parecia contradizer todo o espírito das equações diofantinas.

AQUI VALE A PENA dar alguns detalhes. Mais adiante eles vão nos ajudar um bocado. Não se deixe intimidar pela álgebra; ela está aí basicamente para me dar algo específico a que me referir. Concentre-se na história por trás dela.

As trincas pitagóricas são soluções, em números inteiros, da equação pitagórica

$$x^2 + y^2 = z^2$$

Dividindo por z^2, obtemos

$$\left(\frac{x}{z}\right)^2 + \left(\frac{y}{z}\right)^2 = 1$$

Pelo que vimos no Capítulo 3, isso nos diz que o par de números racionais $(x/z, y/z)$ fica sobre um círculo unitário no plano. Agora, a equação pitagórica originou-se na geometria, e sua interpretação é que o triângulo associado a esses números tem um ângulo reto. A fórmula que acabei de deduzir fornece uma interpretação geométrica ligeiramente diferente, não de apenas uma trinca pitagórica, mas de todas elas. As soluções para a equação pitagórica correspondem diretamente a todos os pontos racionais sobre o círculo unitário. Aqui dizemos que o ponto é racional quando ambas as suas coordenadas o são.

Você pode deduzir uma porção de fatos interessantes dessa inter-relação. Com um pouco de trigonometria, ou diretamente por álgebra, poderá descobrir que para qualquer número t o ponto

$$\left(\frac{2t}{t^2+1}, \frac{t^2-1}{t^2+1}\right)$$

fica sobre um círculo unitário. Mais ainda, se t é racional, o ponto também é. Todos os pontos racionais surgem dessa maneira, logo temos uma fórmula completa para todas as soluções da equação pitagórica. É equivalente à fórmula de Euclides, que é a mesma de Diofanto. Como exemplo, se $t = 22/7$, a fórmula resulta em

$$\left(\frac{308}{533}, \frac{435}{533}\right)$$

e você pode verificar que $308^2 + 435^2 = 533^2$. Para nós, a fórmula precisa não é extremamente importante; o que importa é que exista uma.

Essa não é a única equação diofantina para a qual uma fórmula fornece todas as soluções, mas elas são relativamente raras. Outras incluem as chamadas equações de Pell, tais como $x^2 = 2y^2 + 1$. Esta tem infinitas soluções, tais como $3^2 = 2 \times 2^2 + 1$, $17^2 = 2 \times 12^2 + 1$, e há uma fórmula geral. No entanto, as trincas pitagóricas têm mais estrutura que isso, também derivada da geometria. Suponha que você tenha duas trincas pitagóricas. Então existem duas soluções correspondentes da equação pitagórica – pontos racionais sobre o círculo. A geometria fornece um meio natural de "somar" esses dois pontos. Comece pelo ponto (1,0) no qual o círculo corta o eixo horizontal, e encontre os ângulos entre esse ponto e as duas soluções. Some os dois ângulos (Figura 25) e veja qual é o ponto resultante. Ele certamente fica sobre o círculo. Um breve cálculo mostra que ele é racional. Logo, a partir de duas soluções quaisquer, podemos deduzir uma terceira. Os matemáticos já notaram muitos fatos como esse. A maioria deles faz sentido imediato se você pensar em pontos racionais sobre um círculo.

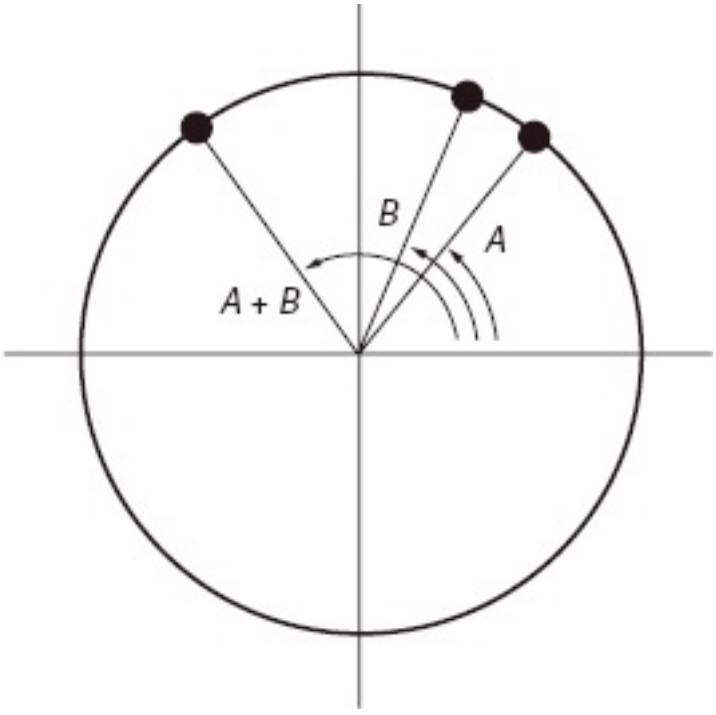


FIGURA 25 Combinação de duas soluções racionais *A* e *B* da equação pitagórica para obter uma terceira, *A + B*.

O “breve cálculo” que introduzi disfarçadamente faz uso da trigonometria. As funções trigonométricas clássicas, tais como o seno e o cosseno, estão intimamente relacionadas com a geometria do círculo. O cálculo ao qual aludi, utiliza fórmulas padronizadas, bem elegantes, para o seno e o cosseno da soma de dois ângulos em termos dos senos e cossenos dos próprios ângulos que estão sendo somados. Há muitas maneiras de se estabelecer senos e cossenos, e uma bem adequada vem do cálculo integral. Se você integrar a função algébrica $1/\sqrt{1-x^2}$, o resultado pode ser expresso em termos da função seno. Na verdade, precisamos é da função inversa do seno: o ângulo cujo seno é o número em que estamos pensando.[2]

A integral surge quando tentamos deduzir uma fórmula para o comprimento do arco de um círculo usando o cálculo, e a geometria do círculo tem uma implicação simples, mas muito importante, para o resultado. O comprimento da circunferência do círculo unitário vale 2π, então percorrer uma distância 2π em torno de um círculo o traz de volta exatamente para o mesmo ponto. A mesma coisa vale para qualquer múltiplo inteiro de 2π: pela convenção matemática padrão, inteiros positivos correspondem ao sentido anti-horário, negativos ao sentido horário. Segue-se daí que o seno e o cosseno de um número permanecem inalterados se for somado um múltiplo inteiro de 2π a esse número. Dizemos que as funções são periódicas, com período 2π.

Os analistas dos séculos XVIII e XIX descobriram uma vasta generalização dessa integral, junto com uma tropa de novas funções interessantes análogas às trigonométricas familiares. Essas novas funções eram intrigantes; eram periódicas, como o seno e o cosseno, mas de forma muito mais elaborada. Em vez de ter um período, como 2π (e seus múltiplos inteiros), tinham dois períodos independentes. Se você tentar fazer isso com funções reais, tudo que obterá são constantes, mas para funções complexas as possibilidades são muito mais ricas.

Essa área foi iniciada pelo matemático italiano Giulio di Fagnano e pelo prolífico Euler. Fagnano estava tentando encontrar o comprimento do arco de uma elipse usando o cálculo, mas não conseguiu achar uma fórmula explícita – o que não é mais surpresa, pois hoje sabemos que não existe. Contudo, ele notou uma relação entre os comprimentos de vários arcos especiais, que publicou em 1750. Euler notou a mesma relação no mesmo contexto, e a apresentou como uma relação formal entre integrais. São

similares à integral associada com a função seno, mas a expressão de segundo grau $1 - x^2$ sob a raiz quadrada é substituída por um polinômio de terceiro ou quarto graus, por exemplo $(1 - x^2)(1 - 4x^2)$.

Em 1811, Adrien-Marie Legendre publicou seu primeiro livro em um pesado tratado em três volumes sobre essas integrais, que são conhecidas como integrais elípticas devido à sua conexão com o comprimento do arco de um segmento de elipse. No entanto, ele deixou passar a característica mais significativa dessas integrais: a existência de novas funções, análogas ao seno e cosseno, cujas funções *inversas* expressam o valor da integral de maneira simples.[3] Gauss, Niels Henrik Abel e Carl Jacobi rapidamente perceberam a negligência. Gauss, como sempre, guardou a descoberta para si. Abel submeteu um artigo à Academia Francesa em 1826, mas Cauchy, o presidente, extraviou o manuscrito, e o artigo só foi publicado em 1841, doze anos após a trágica morte precoce de Abel devido a uma enfermidade pulmonar. Todavia, outro artigo de Abel sobre o mesmo tópico foi publicado em 1827. Jacobi fez dessas novas "funções elípticas" a base para seu enorme tomo, publicado em 1829, que impulsionou a análise complexa rumo a uma trajetória inteiramente nova.

O que emergiu foi um belíssimo pacote de propriedades inter-relacionadas, análogas às das funções trigonométricas. A relação notada por Fagnano e Euler podia ser reinterpretada como uma simples lista de fórmulas relacionando funções elípticas da soma de dois números com as funções elípticas dos próprios números. O traço mais admirável das funções elípticas supera as funções trigonométricas de um jeito espetacular. Não só as funções elípticas são periódicas: elas são duplamente periódicas. Uma linha é unidimensional, então padrões podem se repetir em apenas uma direção ao longo da linha. O plano complexo é bidimensional, então os padrões podem se repetir feito papel de parede: de cima a baixo pela folha e também lateralmente pela parede, em faixas de papel adjacentes. Associados à cada função elíptica há dois números complexos independentes, seus períodos, e somar qualquer um deles com a variável não altera o valor da função.

Repetindo esse processo, concluímos que o valor da função não se altera se somarmos com a variável qualquer combinação inteira dos dois períodos. Essas combinações possuem interpretação geométrica: determinam uma grade no plano complexo. A grade especifica um "ladrilhamento" do plano por meio de paralelogramos, e o que quer que aconteça em um paralelogramo é copiado em todos os outros (Figura 26).

Se considerarmos apenas um paralelogramo, a maneira como ele se liga às cópias adjacentes significa que temos que tornar idênticos os lados opostos, da mesma forma que um toro é definido tornando idênticos os lados opostos de um quadrado (Figura 12). Um paralelogramo com lados opostos idênticos é também um toro topológico. Logo, assim como o seno e o cosseno estão relacionados com o círculo, as funções elípticas estão relacionadas com o toro.

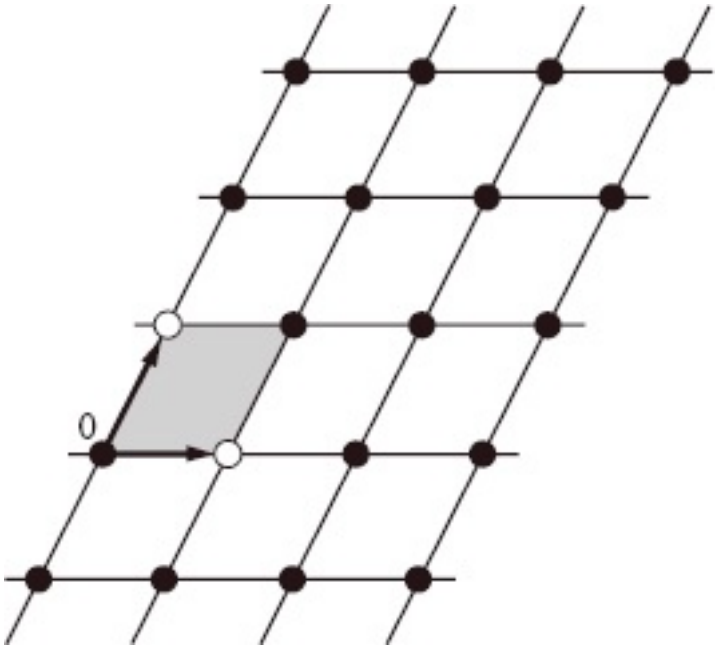


FIGURA 26 Grade no plano complexo. As setas apontam para os dois períodos, mostrados como pontos brancos. O valor da função no paralelogramo sombreado determina o valor em todo outro paralelogramo.

Há também um elo com a teoria dos números. Eu disse que a função inversa do seno é obtida integrando uma fórmula que envolve a raiz quadrada de um polinômio de segundo grau. As funções elípticas são similares, mas o polinômio de segundo grau é substituído por um de terceiro ou quarto graus. O caso quártico foi já mencionado de forma breve, porque historicamente veio antes, mas agora vamos focalizar o caso cúbico. Se representarmos a raiz quadrada por y, e o polinômio por $ax^3 + bx^2 + cx + d$, onde a, b, c e d são coeficientes numéricos, então x e y satisfazem a equação

$$y^2 = ax^3 + bx^2 + cx + d$$

Esta equação pode ser considerada em diversos contextos distintos, dependendo das restrições colocadas nas variáveis e nos coeficientes. Se

forem reais, a equação define uma curva no plano. Se forem complexos, os geômetras algébricos ainda chamam o conjunto de soluções de curva, por analogia. Mas agora é uma curva no espaço de pares de números complexos, que é quadridimensional em coordenadas reais. E a curva é na realidade uma superfície, deste ponto de vista dos números reais.

A Figura 27 mostra as curvas elípticas reais $y^2 = 4x^3 - 3x + 2$ e $y^2 = 4x^3 - 3x$, que são típicas. Como y aparece ao quadrado, a curva é simétrica em relação ao eixo horizontal. Dependendo dos coeficientes, ela é ou uma curva única sinuosa ou tem um componente oval separado. Nos números complexos, a curva é sempre uma peça inteira.

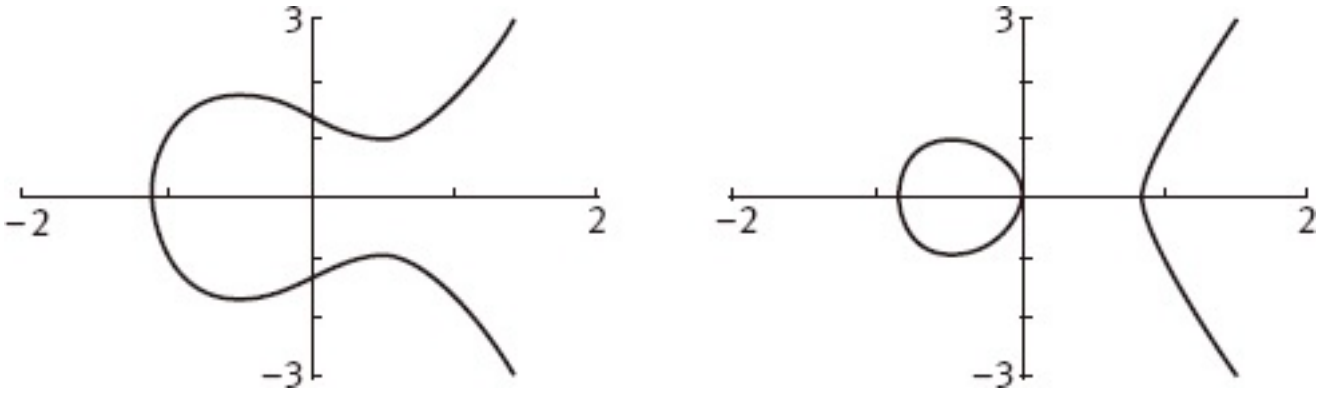


FIGURA 27 Curvas elípticas reais típicas. *Esquerda*: $y^2 = 4x^3 - 3x + 2$. *Direita*: $y^2 = 4x^3 - 3x$.

A teoria dos números entra em jogo quando estabelecemos que as variáveis e os coeficientes sejam racionais. Agora estamos olhando para uma equação diofantina. Ela é confusamente chamada de curva elíptica, embora sua aparência não tenha nada a ver com uma elipse; o motivo é a sua ligação com as funções elípticas. É como chamar um círculo de curva triangular devido à sua ligação com a trigonometria. Infelizmente, o nome está agora inscrito em tabletes de pedra, de modo que temos de conviver com ele.

Como as funções elípticas têm uma teoria rica e profunda, os teóricos dos números descobriram inúmeras e belas propriedades das curvas elípticas. Uma é estreitamente análoga à maneira como podemos combinar duas soluções da equação pitagórica somando os ângulos correlacionados. Dois pontos de uma curva elíptica podem ser combinados desenhando-se uma reta que passe por eles e vendo onde ela cruza a curva pela terceira vez (Figura 28). (Sempre há esse terceiro ponto, porque a equação é cúbica. No entanto, ele poderá estar “no infinito” ou coincidir com um dos dois primeiros pontos caso a reta traçada seja tangente à curva.) Se os dois

pontos são *P* e *Q*, representemos o terceiro por *P* * *Q*.

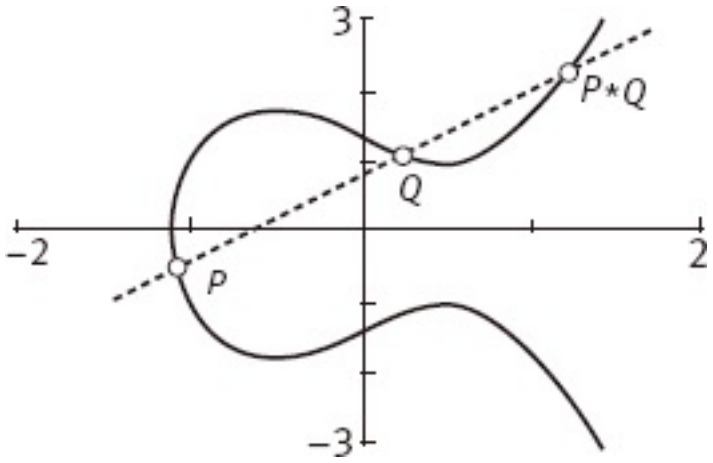


FIGURA 28 Combinação dos pontos *P*, *Q* para obter o ponto *P* * *Q*.

Um cálculo mostra que se *P* e *Q* são pontos racionais, então *P* * *Q* também é. A operação * dá ao conjunto de pontos racionais uma estrutura algébrica, mas revela-se útil considerar uma operação correlacionada. Escolhemos qualquer ponto *O* na curva, e definimos

$$P + Q = (P * Q) * O$$

Essa nova operação obedece a algumas das leis básicas da álgebra comum, com *O* comportando-se como zero, e transforma o conjunto de todos os pontos racionais naquilo que os algebristas chamam de grupo (Capítulo 10). O aspecto essencial é que, como as trincas pitagóricas, você pode "somar" duas soluções quaisquer para obter uma terceira. A ocorrência dessa "lei de grupos" nos pontos racionais é surpreendente, e de modo particular significa que, uma vez tendo encontrado duas soluções racionais da equação diofantina, automaticamente obtemos outras mais.

Por volta de 1908, Poincaré perguntou se existiria uma quantidade finita de soluções a partir das quais todas as outras podem ser obtidas aplicando-se a operação de grupos vezes e vezes repetidas. É um resultado importante porque implica que *todas* as soluções racionais podem ser caracterizadas anotando-se uma lista finita. Em seu espetacular artigo de 1922, Mordell provou que a resposta para a pergunta de Poincaré é "sim". Agora as curvas elípticas adquiriram uma importância central na teoria dos números, porque era incomum ter esse tipo de controle sobre qualquer equação diofantina.

TANTO AS EQUAÇÕES PITAGÓRICAS quanto as curvas elípticas, portanto, têm infinitas soluções racionais. Em contraste, muitas equações diofantinas têm apenas uma quantidade finita de soluções, muitas vezes nenhuma. Vou divagar um pouco e discutir a família inteira de tais equações, e a recente e notável prova de que as soluções óbvias são as únicas que existem.

Os pitagóricos estavam interessados em suas equações porque acreditavam que o universo fundava-se em números. Para sustentar essa filosofia, descobriram que razões numéricas simples governam a harmonia musical. Observaram esse fato experimentalmente usando uma corda esticada. Uma corda sob a mesma tensão que tenha a metade do comprimento emite uma nota uma oitava mais alta. Essa é a combinação mais harmoniosa de duas notas: tão harmoniosa que soa um pouco sem graça. Na música ocidental as harmonias seguintes mais importantes são a quarta, onde uma corda tem $^3/_4$ do comprimento da outra, e a quinta, onde uma corda tem $^2/_3$ do comprimento da outra.[4]

Começando com 1 e multiplicando repetidamente por 2 ou 3 obtemos os números 2, 3, 4, 6, 8, 9, 12, e assim por diante – números da forma 2^a3^b. Devido à ligação com a música esses números ficaram conhecidos como números harmônicos. No século XIII um escritor judeu que vivia na França escreveu *Sha'ar ha-Shamayim* (O portão do céu), uma enciclopédia baseada em fontes árabes e gregas. Ele a dividiu em três partes: física, astronomia e metafísica. Seu nome era Gerson ben Solomon Catalan. Em 1343, o bispo de Meaux persuadiu o filho de Gerson (bem, os historiadores acham que provavelmente foi seu filho) Levi ben Gerson a escrever um livro de matemática, *A harmonia dos números*, incluindo um problema levantado pelo compositor e teórico musical Philippe de Vitry: quando a diferença entre dois números harmônicos pode ser 1? É fácil encontrar tais pares: De Vitry sabia de quatro, isto é (1, 2), (2, 3), (3, 4) e (8, 9). Ben Gerson provou que essas são as únicas soluções possíveis.

Entre os pares harmônicos de De Vitry, o mais interessante é (8, 9). O primeiro é um cubo, 2^3; o segundo é um quadrado, 3^2. Os matemáticos começaram a se perguntar se outros quadrados e cubos poderiam diferir de 1 unidade, e Euler provou que não, além do caso trivial (0, 1), e também (–1, 0), se forem permitidos números negativos. Em 1844, o segundo Catalan da história fez imprimir uma alegação mais abrangente, na qual muitos matemáticos devem ter pensado mas não se deram ao trabalho de explicitar. Era o matemático belga Eugène Charles Catalan, que escreveu a uma das principais publicações matemáticas da época, o *Journal für die*

Reine und Angewandte Mathematik (Revista da matemática pura e aplicada):

Eu rogo, senhor, que por favor anuncie em sua publicação o seguinte teorema que acredito ser verdadeiro embora não tenha ainda conseguido prová-lo completamente; talvez outros tenham mais êxito. Dois números inteiros consecutivos, diferentes de 8 e 9, não podem ser potências consecutivas; dito de outra maneira, a equação $x^m - y^n = 1$, na qual as incógnitas são inteiros positivos admite apenas uma solução.

Essa afirmação tornou-se conhecida como conjectura de Catalan. Os expoentes *m* e *n* são inteiros maiores que 1.

Apesar do progresso parcial, a conjectura de Catalan recusava-se teimosamente a render-se, até que foi provada em 2002 por Preda Mihăilescu. Nascido na Romênia em 1955, estabelecera-se na Suíça em 1973, e apenas recentemente completara seu doutorado. O título de sua tese era "Ciclotomia de anéis e teste de primalidade", e aplicava a teoria dos números ao teste de primalidade (Capítulo 2). Esse problema não tinha relação particular alguma com a conjectura de Catalan, mas Mihăilescu veio a perceber que seus métodos com certeza tinham. Eles derivavam das ideias que mencionei no Capítulo 3: a construção de Gauss do heptadecágono, o polígono de dezessete lados e equações algébricas correlatas, cujas soluções são chamadas números ciclotômicos. A prova era altamente técnica e veio como um choque para a comunidade matemática. Ela nos diz que quaisquer que sejam os valores que escolhamos para as duas potências, o número de soluções é finito – e à parte as soluções óbvias usando 0 e ±1, a única interessante é $3^2 - 2^3 = 1$.

OS EXEMPLOS ACIMA mostram que algumas equações diofantinas têm infinitas soluções, outras não. Grande coisa – essas alternativas cobrem tudo. Porém, se você começa a perguntar quais equações são de que tipo, a coisa fica mais interessante. Mordell, perito em equações diofantinas, estava escrevendo um livro seminal. Em sua época, a área parecia como os primórdios da biologia: um monte de gente caçando borboletas e muito pouco em termos de classificação sistemática. Uma Pitagórica Pintada aqui, uma Grande Azul Elíptica ali, e no mato lagartas da Pelliana Malhada. O campo estava exatamente como Diofanto o deixara, só que mais: uma lista

sem estrutura de recursos separados, um para cada tipo de equação. Isso é material pobre para um livro-texto, e necessitava desesperadoramente de organização, então Mordell propôs-se a fazer exatamente isso.

Em algum momento ele deve ter notado que todas as equações que sabemos ter infinitas soluções – tais como a equação pitagórica e as curvas elípticas – possuíam um fator comum. Ele se concentrou em uma classe de equações, aquelas que (após serem convertidas em equações de números racionais, como fez com Pitágoras) envolvem apenas duas variáveis. Uma é exemplificada pela equação pitagórica na forma equivalente $x^2 + y^2 = 1$. Aqui há uma fórmula para as soluções. Ligue qualquer número racional à fórmula e você obtém uma solução racional, e surgem todas as soluções. A outra é exemplificada pelas curvas elípticas: há um *processo* que gera novas soluções a partir de antigas, e uma garantia de que, se você começar com um conjunto finito adequado de soluções, o processo as produz todas.

A conjectura de Mordell afirma que sempre que houver infinitas soluções racionais, deve-se aplicar um desses dois recursos. Ou existe uma fórmula geral ou existe um processo que gera todas as soluções a partir de um conjunto finito adequado delas. Em todos os outros casos, o número de soluções racionais é finito, por exemplo nas equações $x^m - y^n = 1$ que aparecem na conjectura de Catalan. Em certo sentido, as soluções são então apenas coincidências, sem nenhuma estrutura subjacente.

Mordell chegou a essa observação de maneira ligeiramente distinta. Ele notou que toda equação com infinitas soluções racionais tem uma surpreendente característica topológica. Possui genus 0 ou 1. Recorde-se do Capítulo 4 que genus é um conceito da topologia de superfícies, e conta quantos furos a superfície tem. Uma esfera tem genus 0, um toro genus 1, um toro com dois furos genus 2, e assim por diante. Como é que superfícies entram em um problema de teoria dos números? A partir da geometria de coordenadas. Vimos que a equação pitagórica, interpretada em termos de números racionais e expandida de modo a permitir números reais como soluções, determina um círculo. Mordell deu um passo adiante, e permitiu números complexos como soluções. Qualquer equação em duas variáveis complexas determina aquilo que o geômetra algébrico chama de curva complexa. No entanto, do ponto de vista dos números reais e do sistema visual humano, todo número complexo é bidimensional: tem dois componentes reais, suas partes real e imaginária. Assim, a "curva" para olhos complexos é uma superfície para você e para mim. Sendo uma superfície, tem genus: e lá vamos nós.

Sempre que uma equação era conhecida como tendo apenas uma quantidade finita de soluções, seu genus era pelo menos 2. Equações importantes cujo status era desconhecido também tinham genus pelo menos 2. Num salto bravio e destemido, baseado no que parecia ser, à época, uma evidência bastante efêmera, Mordell conjecturou que qualquer equação diofantina de genus 2 ou mais tem apenas uma quantidade finita de soluções racionais. De um só golpe, as borboletas de Diofanto foram de maneira ordenada arranjadas em famílias correlacionadas; apropriadamente, por genus.

Havia apenas um minúsculo senão na conjectura de Mordell. Ela relacionava duas coisas absolutamente diferentes: soluções racionais e topologia. Na época, qualquer elo plausível era tênue ao extremo. Se existisse alguma ligação, ninguém sabia como fazer para encontrá-la. Logo, era uma conjectura selvagem, especulação não substanciada, mas o retorno potencial era gigantesco.

Em 1983, Faltings publicou uma prova drástica de que a selvagem especulação de Mordell de fato estava certa. Sua prova fez uso de métodos profundos da geometria algébrica. Uma prova muito diferente, baseada em aproximar números reais de números racionais, foi achada por Paul Vojta, e, em 1990, Enrico Bombieri publicou uma prova simplificada nas mesmas linhas. Há uma aplicação do teorema de Faltings ao último teorema de Fermat, problema esse do qual trataremos extensivamente no Capítulo 7. Ele afirma que para qualquer inteiro n maior ou igual a 3, a equação $x^n + y^n = 1$ tem apenas uma quantidade finita de soluções. O genus da curva associada é $(n-1)(n-2)/2$, e isso será pelo menos 3 se n for 4 ou mais. O teorema de Faltings implica de imediato que para qualquer $n \geq 4$, a equação de Fermat tem, no máximo, uma quantidade finita de soluções racionais. Fermat alegou que não havia nenhuma exceto quando x ou y é zero, então isso já foi um grande avanço. No próximo capítulo, atacamos a história do último teorema de Fermat, e vamos ver como a alegação de Fermat foi plenamente justificada.

7. Margens inadequadas

O último teorema de Fermat

ENCONTRAMOS FERMAT pela primeira vez no Capítulo 2, onde seu elegante teorema sobre potências de números forneceu um método para se verificar se os números são primos. Este capítulo fala sobre uma afirmação muito mais difícil: o último teorema de Fermat. Soa muito misterioso. "Teorema" parece claro, mas quem foi Fermat, e por que foi seu *último* teorema? Seria o nome uma astuta jogada de marketing? Não; o nome ficou atado ao problema no século XVIII, quando apenas alguns matemáticos de destaque tinham ouvido falar ou se importavam com ele. Mas o último teorema de Fermat realmente é misterioso.

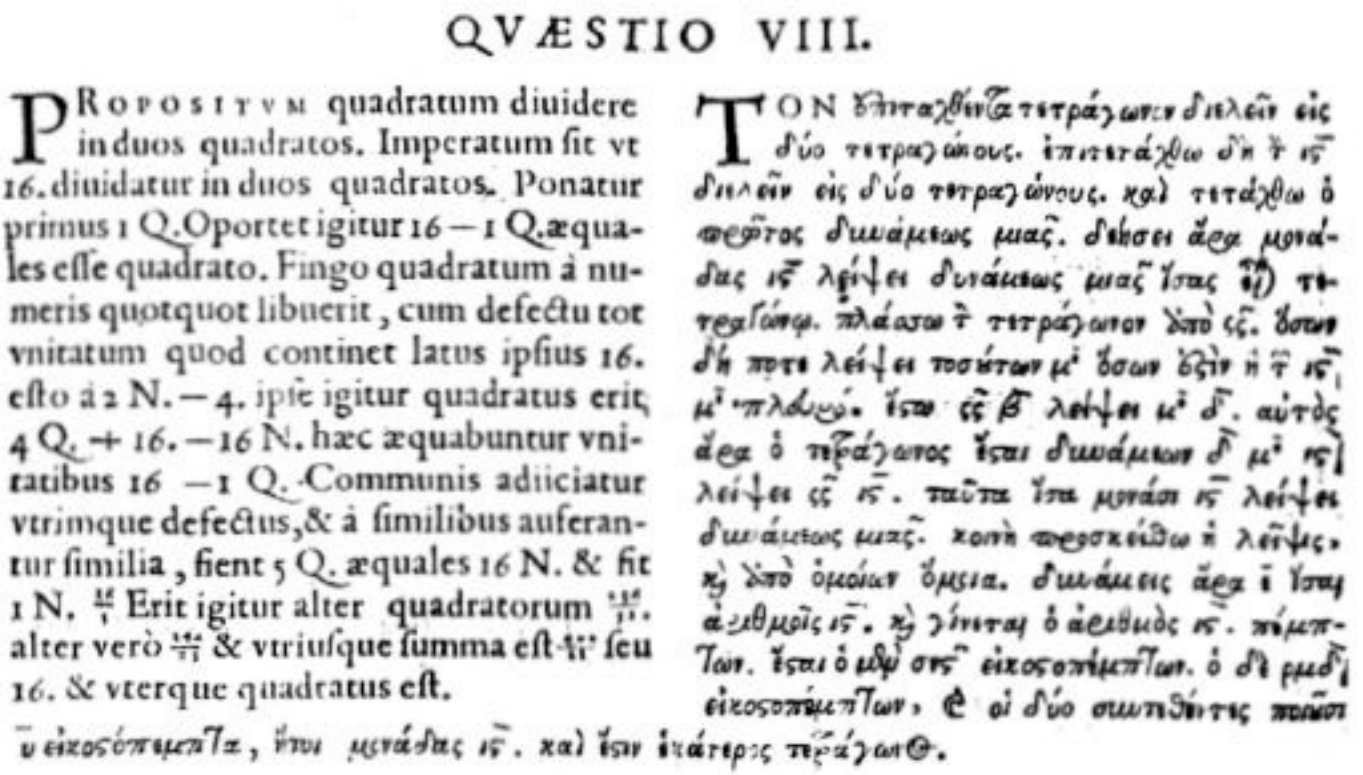
Segundo algumas fontes, Pierre Fermat teria nascido na França em 1601, porém, segundo outras, teria sido em torno de 1607-8. A discrepância talvez venha da confusão com um irmão de mesmo nome. Seu pai era um bem-sucedido mercador de couro e detinha uma alta posição no governo local, e sua mãe provinha de uma família de advogados parlamentares. Ele foi para a Universidade de Toulouse, mudou-se para Bordeaux no fim da década de 1620 e ali mostrou sinais promissores de talento matemático. Era fluente em diversos idiomas e empreendeu a restauração de uma obra perdida de matemática grega clássica da autoria de Apolônio. Compartilhou suas muitas descobertas com os mais importantes matemáticos de seu tempo.

Em 1631, tendo se diplomado em direito na Universidade de Orléans, foi nomeado conselheiro da Alta Corte de Judicatura em Toulouse. Isso lhe autorizava a mudar seu sobrenome para "de Fermat", mantendo-se conselheiro para o resto da vida. Sua paixão, porém, era a matemática. Publicou pouco, preferindo escrever cartas explicando suas descobertas, geralmente sem prova. Seu trabalho recebeu o devido reconhecimento dos profissionais, muitos dos quais tratava pelo primeiro nome, enquanto mantinha seu status de amador. Mas Fermat possuía tanto talento que na verdade era profissional; apenas não detinha uma posição oficial em matemática.

Algumas de suas provas sobreviveram em cartas e artigos, e fica claro

que Fermat sabia o que era uma prova genuína. Depois da sua morte, muitos dos seus teoremas mais profundos permaneceram não provados, e os profissionais mergulharam no trabalho com eles. Em algumas décadas, apenas uma das afirmações de Fermat ainda carecia de prova, então, naturalmente, passou a ser conhecida como seu último teorema. Ao contrário dos outros, ele não sucumbiu, e logo se tornou notório pelo contraste entre a facilidade da afirmação e a aparente dificuldade de achar uma prova.

Fermat parece ter conjecturado seu último teorema por volta de 1630. A data exata é desconhecida, mas foi então que Fermat começou a ler uma edição recentemente publicada da *Arithmetica* de Diofanto. E foi daí que tirou a ideia. O último teorema foi impresso pela primeira vez em 1670, cinco anos após a morte de Fermat, quando seu filho Samuel publicou uma edição da *Arithmetica*. Essa edição tinha uma novidade interessante. Incorporou as anotações que Pierre escrevera nas margens de seu exemplar pessoal da tradução latina de 1621 feita por Claude Gaspard Bachet de Méziriac. O último teorema é afirmado como uma nota anexa à questão VIII do Livro II de Diofanto (Figura 29).



QVÆSTIO VIII.

PROPOSITVM quadratum diuidere in duos quadratos. Imperatum sit vt 16. diuidatur in duos quadratos. Ponatur primus 1 Q. Oportet igitur 16 − 1 Q. æquales esse quadrato. Fingo quadratum à numeris quotquot libuerit, cum defectu tot vnitatum quod continet latus ipsius 16. esto à 2 N. − 4. ipse igitur quadratus erit 4 Q. + 16. − 16 N. hæc æquabuntur vnitatibus 16 − 1 Q. Communis adiiciatur vtrimque defectus, & à similibus auferantur similia, fient 5 Q. æquales 16 N. & fit 1 N. $\frac{16}{5}$. Erit igitur alter quadratorum $\frac{256}{25}$. alter verò $\frac{144}{25}$ & vtriusque summa est $\frac{400}{25}$ seu 16. & vterque quadratus est.

Τὸν δοθέντα τετράγωνον διελεῖν εἰς δύο τετραγώνους. ἐπιτετάχθω δὴ τὸν ιϛ̄ διελεῖν εἰς δύο τετραγώνους. καὶ τετάχθω ὁ ἑνὸς δυνάμεως μιᾶς. δεήσει ἄρα μονάδας ιϛ̄ λεῖψει δυνάμεως μιᾶς ἴσας τῷ) τετραγώνῳ. πλάσω τὸ τετράγωνον ἀπὸ ϛ̄ς. ὅσων δὴ ποτε λείψει τοσούτων μ̄ ὅσων ἔχει ἡ τοῦ ιϛ̄ μ̄ πλευρά. ἔστω ϛ̄ς β̄ λείψει μ̄ δ̄. αὐτὸς ἄρα ὁ τετράγωνος ἔσται δυνάμεων δ̄ μ̄ ιϛ̄ λείψει ϛ̄ς ιϛ̄. ταῦτα ἴσα μονάσι ιϛ̄ λείψει δυνάμεως μιᾶς. κοινὴ προσκείσθω ἡ λεῖψις. καὶ ἀπὸ ὁμοίων ὁμοῖα. δυνάμεις ἄρα ε̄ ἴσαι ἀριθμοῖς ιϛ̄. καὶ γίνεται ὁ ἀριθμὸς ιϛ̄. πέμπτων. ἔσται ὁ μὲν τοῦ εἰκοσοπέμπτων. ὁ δὲ μδ̄ εἰκοσοπέμπτων, ἑκάτερος τετράγωνος.

ὁ δὲ ιϛ̄ εἰκοσόπεμπτα, ἴσα μονάδας ιϛ̄. καὶ ἔστιν ἑκάτερος τετράγωνος.

OBSERVATIO DOMINI PETRI DE FERMAT.

CVbum autem in duos cubos, aut quadratoquadratum in duos quadratoquadratos & generaliter nullam in infinitum vltra quadratum potestatem in duos eiusdem nominis fas est diuidere cuius rei demonstrationem mirabilem sane detexi. Hanc marginis exiguitas non caperet.

FIGURA 29 A nota de Fermat escrita na margem, publicada na edição de seu filho da *Arithmetica* de Diofanto.

O problema ali resolvido é escrever um quadrado perfeito como soma de dois quadrados perfeitos. No Capítulo 6 vimos que existem infinitas dessas trincas pitagóricas. Diofanto faz uma pergunta correlacionada, porém mais difícil: como achar os dois lados menores de um triângulo, dado o lado maior? Um quadrado específico precisa ser "dividido" em dois quadrados, ou seja, expressos como sua soma. Ele mostra como solucionar esse problema quando o lado maior do triângulo é 4, obtendo a resposta

$$4^2 = \left(\frac{16}{5}\right)^2 + \left(\frac{12}{5}\right)^2$$

em números racionais. Multiplicando por 25 obtemos $20^2 = 16^2 + 12^2$ e dividindo por 16 tem-se o familiar $3^2 + 4^2 = 5^2$. Era típico de Diofanto ilustrar métodos gerais com exemplos específicos, uma tradição que

remonta à antiga Babilônia, e não fornecia provas.

O exemplar pessoal de Fermat da *Arithmetica* não sobreviveu, mas ele deve ter escrito nele sua nota marginal, porque Samuel assim o afirma. É improvável que Fermat tivesse deixado um tesouro desses oculto por muito tempo, e sua conjectura é tão natural que ele provavelmente pensou nela assim que leu a questão VIII do Livro II. Evidentemente, perguntou-se se algo similar poderia ser conseguido usando cubos em vez de quadrados, uma pergunta natural para um matemático. Não encontrou exemplos – podemos ter certeza disso, pois não há nenhum – e do mesmo modo não obteve sucesso quando tentou potências mais altas, por exemplo, a quarta potência. Concluiu que essas questões não tinham soluções. Sua nota na margem da página diz muito:

> É impossível dividir um cubo em dois cubos, ou uma quarta potência em suas quartas potências, ou em geral qualquer potência maior que a segunda em duas potências idênticas. Descobri uma prova verdadeiramente maravilhosa disso, mas essa margem é estreita demais para contê-la.

Em linguagem algébrica, Fermat alegava que a equação diofantina

$$x^n + y^n = z^n$$

não tem solução com números inteiros se n for qualquer inteiro igual ou maior que 3. É claro que ele estava ignorando soluções triviais nas quais x ou y é zero. Para evitar repetir a fórmula por toda parte, vou referir-me a ela como a equação de Fermat.

Se Fermat realmente tinha uma prova, ninguém jamais a encontrou. O teorema foi finalmente provado em 1995, mais de três séculos e meio depois de ter sido afirmado, mas os métodos vão muito além de qualquer coisa que ele tivesse à disposição em sua época, ou que pudesse ter inventado. A busca por uma prova teve enorme influência no desenvolvimento da matemática. Foi praticamente a causa da criação da teoria algébrica dos números, que floresceu no século XIX devido a uma tentativa fracassada de provar o teorema e uma ideia brilhante que parcialmente o resgatou. No fim do século XX e no século XXI, deflagrou uma revolução.

OS PRIMEIROS A TRABALHAR no último teorema de Fermat tentaram pegar as potências uma a uma. A prova geral de Fermat, aludida na margem, pode ou não ter existido, mas sabemos como ele provou o teorema para a quarta potência. A principal ferramenta foi a receita de Euclides de trincas pitagóricas. A quarta potência de qualquer número é o quadrado do quadrado desse número. Logo, qualquer solução da equação de Fermat para a quarta potência é uma trinca pitagórica, para a qual todos os números são eles mesmos quadrados. Essa condição adicional pode ser inserida na receita de Euclides e, após algumas manobras astutas, o que emerge é *outra* solução para a equação de Fermat para a quarta potência.[1] Isso pode parecer que não é progresso; depois de uma página de álgebra o problema é reduzido ao mesmo problema. Contudo, ele fica realmente reduzido: os números na segunda solução são menores que os da primeira, hipotética. Crucialmente, se a primeira solução não é trivial – se x e y são diferentes de zero – então o mesmo vale para a segunda solução. Fermat mostrou que a repetição desse procedimento levaria a uma sequência de soluções nas quais os números se tornariam perpetuamente menores. No entanto, qualquer sequência decrescente de números inteiros precisa cessar. Essa é uma contradição lógica, então a solução hipotética não existe. Ele chamou esse método de "declínio infinito". Agora o reconhecemos como uma prova por indução matemática, mencionada no Capítulo 4, e pode ser reformulada em termos de contraexemplos mínimos. Ou, nesse caso, de modelos de virtude mínimos. Suponha que exista um cidadão virtuoso, uma solução não trivial para a equação. Então existe um cidadão virtuoso mínimo. Mas o argumento de Fermat implica a existência e um cidadão virtuoso ainda menor – contradição. Portanto nenhum cidadão pode ser virtuoso. Diferentes provas para a quarta potência vêm aparecendo desde então, e são conhecidas cerca de trinta.

Fermat explorou o simples fato de que a quarta potência é um tipo especial de quadrado. A mesma ideia mostra que, para provar o último teorema de Fermat, pode-se assumir que a potência n é ou 4 ou um primo ímpar. Qualquer número n maior que dois é divisível por 4 ou por um primo ímpar p, então toda enésima potência é ou uma quarta potência ou uma potência p. Nos dois séculos seguintes, o último teorema de Fermat foi provado para exatamente três primos ímpares: 3, 5 e 7. Euler lidou com cubos em 1770; embora haja uma lacuna na prova publicada, ela pode ser tapada usando um resultado que Euler publicou em outro lugar. Legendre e Peter Lejeune-Dirichlet lidaram com as quintas potências por volta de

1825. Gabriel Lamé provou o último teorema de Fermat para as sétimas potências em 1839. Muitas provas diferentes foram posteriormente encontradas para esses casos. Ao longo do caminho, vários matemáticos desenvolveram provas quando a potência é 6, 10 e 14, mas eram provas subsidiadas pelas provas para 3, 5 e 7.

Cada prova faz extensivo uso de características algébricas que são especiais para a potência considerada. Não havia indício de qualquer estrutura geral que pudesse provar o teorema para todas as potências, ou mesmo para uma quantidade significativa de potências diferentes. À medida que as potências cresciam, as provas iam se tornando mais e mais complicadas. Eram necessárias ideias novas, e elas precisavam desbravar novos territórios. Sophie Germain, uma das mais importantes matemáticas, dividiu o último teorema de Fermat para uma potência prima p em dois subcasos. No primeiro, nenhum dos números x, y, z é divisível por p. No segundo, um deles é. Considerando primos "auxiliares" especiais relacionados com p, ela provou que o primeiro caso do último teorema de Fermat não tem soluções para uma potência prima ímpar menor que 100. Todavia, foi difícil provar muita coisa em relação a primos auxiliares em geral.

Germain correspondia-se com Gauss, primeiro usando um pseudônimo masculino, e ele ficou muito impressionado com sua originalidade. Quando revelou que era mulher, ficou mais impressionado ainda, e o disse abertamente. Ao contrário de muitos de seus contemporâneos, Gauss não assumia que mulheres eram incapazes de altas conquistas intelectuais, de forma particular pesquisa matemática. Mais tarde, Germain fez uma tentativa sem êxito de provar o primeiro caso do último teorema de Fermat para todas as potências pares, onde mais uma vez é possível explorar a caracterização euclidiana das trincas pitagóricas. Guy Terjanian finalmente livrou-se das potências pares em 1977. O segundo caso parecia uma noz com a casca muito mais dura, e ninguém foi muito longe na tentativa de quebrá-la.

EM 1847 LAMÉ, seguindo adiante a partir de sua prova para a sétima potência, teve uma magnífica ideia. Ela exigia a introdução de números complexos, mas a essa altura todo mundo já vivia feliz com eles. O ingrediente vital era o mesmo que Gauss havia explorado para construir o polígono regular de dezessete lados (Capítulo 3). Todo teórico dos números

o conhecia, mas até Lamé ninguém se perguntara com seriedade se não poderia ser justamente essa a tarefa para provar o último teorema de Fermat.

No sistema dos números reais, 1 tem exatamente uma raiz de ordem p (quando p é ímpar), ou seja, ele mesmo. Mas nos números complexos, 1 tem muitas raízes de ordem p, na verdade exatamente p raízes. Esse fato é uma consequência do teorema fundamental da álgebra, pois essas raízes satisfazem a equação $x^p - 1 = 0$, que tem grau p. Há uma bela fórmula para essas raízes complexas de ordem p para a unidade, como são chamadas, e ela mostra que são as potências $1, \zeta, \zeta^2, \zeta^3 \ldots, \zeta^{p-1}$ de um número complexo específico ζ.[2] A propriedade que define esses números implica que $x^p + y^p$ divide-se em p fatores:

$$x^p + y^p = (x + y)(x + \zeta y)(x + \zeta^2 y) \ldots (x + \zeta^{p-1} y)$$

Pela equação de Fermat, essa expressão também é igual a z^p, que é a potência p de um número inteiro. Agora, é fácil ver que se um produto de números, sem nenhum fator comum, é uma potência p, então cada número é ele próprio uma potência p. Logo, sem considerar alguns pequenos detalhes técnicos, Lamé podia escrever cada fator como uma potência p. Daí ele deduziu uma contradição.

Lamé anunciou a prova resultante do último teorema de Fermat para a Academia de Paris em março de 1847, dando crédito pela ideia básica a Joseph Liouville. Este agradeceu a Lamé, mas indicou uma questão potencial. A afirmação crucial implicando que cada fator é uma potência p não é ponto pacífico. Depende de que a fatoração em primos seja única – não só para inteiros comuns, onde a propriedade é verdadeira, mas para os novos tipos de número que Lamé havia introduzido. Essas combinações de potências de ζ são chamadas inteiros ciclotômicos; a palavra significa "que cortam o círculo", referindo-se à conexão que Gauss tinha explorado. Não só a propriedade da fatoração única em primos não era provada para os inteiros ciclotômicos, disse Liouville, como podia ser falsa.

Outros já tinham dúvidas. Três anos antes, em uma carta, Gotthold Eisenstein escreveu:

Se existisse o teorema que afirma que o produto de dois números complexos pode ser divisível por um número primo apenas quando um dos fatores é divisível – o que parece totalmente óbvio –, então se teria a teoria

inteira [dos números algébricos] num único golpe; mas este teorema é absolutamente falso.

O teorema ao qual ele alude é o passo principal necessário para uma prova de que a fatoração em primos é única. Eisenstein referia-se não só aos números de que Lamé necessitava, mas a números semelhantes surgindo a partir de outras equações, que são chamados números algébricos. Um número algébrico é um número complexo que satisfaz uma equação polinomial com coeficientes racionais. Um inteiro algébrico é um número complexo que satisfaz uma equação polinomial com coeficientes inteiros, na condição de que o coeficiente da potência mais alta de x seja 1. Para cada polinômio desses, obtemos um corpo numérico algébrico associado (significando que podemos somar, subtrair, multiplicar e dividir esses números para obter números do mesmo tipo) e seu anel (mesma coisa, omitindo "dividir") de inteiros algébricos. Esses são os objetos básicos que são estudados na teoria dos números algébricos.

Se, por exemplo, o polinômio é $x^2 - 2$, então ele tem solução $\sqrt{2}$. O corpo consiste em todos os números $a + b\sqrt{2}$, com a e b racionais; o anel de inteiros consiste em todos os números dessa forma com a e b inteiros. Novamente os fatores primos podem ser definidos e são únicos. Há algumas surpresas: o polinômio $x^2 + x - 1$ tem uma solução $(\sqrt{5} - 1)/2$, assim, apesar da fração, trata-se de um algébrico *inteiro*.

Em teoria dos números algébricos, a dificuldade não é definir fatores. Por exemplo, um inteiro ciclotômico é fator de (isto é, divide) outro se o segundo for igual ao primeiro multiplicado por algum inteiro ciclotômico. A dificuldade não é definir primos: um inteiro ciclotômico é primo se não tem fatores, além das "unidades" triviais, que são os inteiros ciclotômicos que dividem 1. Não há problema em relação a decompor um inteiro ciclotômico, ou qualquer outra parte de um número algébrico, em fatores primos. Basta continuar a fatorá-lo até se acabarem os fatores. Há um meio simples de provar que o procedimento se interrompe, e quando isso acontece, todo fator deve ser primo. Então, qual é a dificuldade? A fatoração única. Se você repetir o procedimento, fazendo escolhas diferentes ao longo do caminho, poderá acabar com uma lista diferente de primos.

À primeira vista, é difícil ver como isso acontece. Os fatores primos são os menores pedaços possíveis nos quais um número pode se dividir. É como pegar um brinquedo feito de Lego e separá-lo em seus bloquinhos

componentes. Se houvesse outro jeito de fazer isso, no final um desses blocos poderia ser dividido em dois ou mais. Mas então não seria um bloco. Infelizmente a analogia com o Lego é enganosa. Os números algébricos não são assim. São mais como blocos com emendas móveis, possíveis de serem ligadas de diferentes formas. Divida um bloco de uma maneira específica, e as peças resultantes se fixam, e não podem mais ser separadas. Divida-o de um jeito diferente, e as peças resultantes voltam a se fixar. Mas dessa vez são diferentes.

Vou lhes dar dois exemplos. O primeiro usa apenas inteiros comuns; é fácil de entender, mas possui algumas características não representativas. Depois lhes mostrarei um exemplo autêntico.

Suponha que vivêssemos num universo em que os únicos números existentes fossem 1, 5, 9, 13, 17, 21, 25, e assim por diante – números que no nosso universo real teriam a forma $4k + 1$. Se você multiplicar dois desses números entre si, obterá outro número do mesmo tipo. Definimos tal número como "primo" se não for o produto de dois números menores *desse tipo*. Por exemplo, 25 não é primo porque é 5 × 5, e 5 é um número da lista. Mas 21 *é* primo, porque, nesse novo sentido, seus fatores comuns 3 e 7 não estão na lista. São da forma $4k + 3$, não $4k + 1$. É fácil ver que todo número do tipo especificado é produto de primos nesse novo sentido. A razão é porque os fatores, se existirem, precisam ficar menores. Em último caso, o processo de fatoração precisa parar. E quando para, os fatores envolvidos são primos.

Todavia, essa espécie de fatoração em primos não é única. Considere o número 4.389, que é 4 × 1.097 + 1, portanto da forma requerida. Eis aqui três fatorações distintas em números da forma requerida:

$$4.389 = 21 \times 209 = 33 \times 133 = 57 \times 77$$

Eu digo que, com a nossa atual definição, todos esses fatores são primos. Por exemplo, 57 é primo, porque seus fatores habituais 3 e 19 não são da forma requerida. O mesmo vale para 21, 33, 77, 133 e 209. Agora podemos explicar essa falta de singularidade. Em inteiros comuns

$$4.389 = 3 \times 7 \times 11 \times 19$$

e todos esses fatores têm a forma "errada" $4k + 3$. As três fatorações diferentes em primos, nesse sentido, surgem agrupando esses números em pares:

$$(3 \times 7) \times (11 \times 19)\ (3 \times 11) \times (7 \times 19)\ (3 \times 19) \times (7 \times 11)$$

Precisamos usar pares porque dois números da forma $4k + 3$, multiplicados entre si, produzem um número da forma $4k + 1$.

Esse exemplo mostra que o argumento "os fatores devem ser únicos porque são os menores pedaços" não funciona. É verdade que existem pedaços *ainda menores* ($21 = 3 \times 7$, por exemplo), mas estes não estão no referido sistema. A principal razão para esse exemplo não ser totalmente representativo é que, embora multiplicar números da forma $4k + 1$ produza números da mesma forma, isso não vale para a adição. Por exemplo, $5 + 5 = 10$ não é da forma requerida. Assim, no jargão da álgebra abstrata, não estamos trabalhando num anel.

O segundo exemplo não tem esse defeito, mas em compensação é um pouco mais difícil de analisar. É o anel de inteiros algébricos para o polinômio $x^2 - 15$. O anel consiste em todos os números $a + b\sqrt{15}$, onde a e b são inteiros. Nele, o número 10 tem duas fatorações distintas:

$$10 = 2 \times 5 = (5 + \sqrt{15}) \times (5 - \sqrt{15})$$

Todos os quatro fatores, 2, 5, $5 + \sqrt{15}$ e $5 - \sqrt{15}$ podem ser provados primos.[3]

Tudo isso está muito mais claro agora do que em 1847, mas não demorou muito para mostrar que as dúvidas de Liouville eram justificadas. Uma quinzena após tê-las expressado, Wantzel informou à academia que a singularidade era verdadeira para alguns valores pequenos de p, mas seu método falhava para a 23ª potência. Pouco tempo depois, Liouville disse à academia que a fatoração única em primos é *falsa* para inteiros ciclotômicos correspondentes a $p = 23$. Ernst Kummer havia descoberto esse fato havia três anos, mas não contara a ninguém porque estava trabalhando num método para contornar o obstáculo. A prova de Lamé funcionava para valores menores que p, inclusive alguns novos: 11, 13, 17, 19. Mas para o caso geral, a prova estava em frangalhos. Era uma aula de como não assumir que afirmações matemáticas plausíveis são óbvias. Elas podem nem mesmo ser verdadeiras.

KUMMER VINHA PENSANDO a respeito do último teorema de Fermat em linhas semelhantes às de Lamé. Notou o potencial obstáculo, levou-o a sério, investigou e descobriu que ele derrubava a abordagem. Achou um exemplo explícito de uma fatoração em primos não única para inteiros ciclotômicos

baseada nas raízes de 23ª ordem da unidade. Mas Kummer não era de desistir facilmente, e achou um meio de contornar o obstáculo – ou, no mínimo, mitigar seus piores efeitos. Sua ideia é especialmente transparente no caso dos números tipo $4k + 1$. O modo de restaurar a fatoração única é introduzir alguns números *novos*, fora do sistema em que estamos interessados. Pelo exemplo, o que precisamos são aqueles números ausentes do tipo $4k + 3$. Ou podemos pegar todo o estoque e lançar mão também dos inteiros pares; então teremos os inteiros, que são fechados em termos de adição e multiplicação. Isto é, se você soma ou multiplica dois inteiros, o resultado é um inteiro.

Kummer surgiu com uma versão da mesma ideia. Por exemplo, podemos restaurar a fatoração única em números primos no anel de todos os números $a + b\sqrt{15}$ introduzindo um número novo, ou seja, $\sqrt{5}$. Para obter um anel, descobrimos que precisamos introduzir também $\sqrt{3}$. Agora

$$2 = (\sqrt{5} + \sqrt{3}) \times (\sqrt{5} - \sqrt{3}) \qquad 5 = \sqrt{5} \times \sqrt{5}$$

e

$$5 + \sqrt{15} = \sqrt{5} \times (\sqrt{5} + \sqrt{3}) \qquad 5 - \sqrt{15} = \sqrt{5} \times (\sqrt{5} - \sqrt{3})$$

De modo que ambas as fatorações surgem agrupando-se os quatro números $\sqrt{5}$, $\sqrt{5}$, $\sqrt{5} + \sqrt{3}$, $\sqrt{5} - \sqrt{3}$ de duas maneiras diferentes.

Kummer chamou a esses novos fatores de números ideais, porque em sua formulação geral não eram exatamente números. Eram símbolos que se comportavam um bocado como números. Ele provou que todo inteiro ciclotômico pode ter uma fatoração única em números primos ideais. Foi uma montagem sutil: nem os inteiros ciclotômicos nem os números ideais tinham uma fatoração única em primos. Mas ao usar os números ideais como ingredientes para uma fatoração em primos dos inteiros ciclotômicos, o resultado era uma fatoração única.

Mais tarde, Richard Dedekind descobriu uma reinterpretação mais civilizada do procedimento de Kummer, e é esta que atualmente usamos. Para cada número ideal fora do referido anel, ele associou um *conjunto* de números dentro do anel. E chamou esse conjunto de ideal. Todo número no anel define um ideal: ele consiste em todos os múltiplos daquele número. Se a fatoração em primos é única, todo ideal também é. Quando não é única, há ideais adicionais. Podemos definir produto e soma de ideais e ideais primos, e Dedekind provou que a fatoração *de ideais* em primos é única para todos os anéis de inteiros algébricos. Isso sugere que na maioria dos

problemas deve-se trabalhar com ideais, e não com os números algébricos em si. É claro que isso introduz novas complexidades, mas a alternativa geralmente é estagnar.

Kummer foi capaz de trabalhar com seus números ideais – suficientemente bem para provar uma versão do último teorema de Fermat com algumas hipóteses extras. Mas outros mortais julgaram os números ideais bastante difíceis, se não um tanto místicos. Contudo, uma vez encarado do ponto de vista de Dedekind, os números ideais faziam perfeito sentido, e a teoria dos números algébricos decolou. Uma ideia importante que surgiu foi uma maneira de medir quanto uma fatoração única fracassa em um anel de inteiros algébricos. A cada anel desses corresponde um número inteiro chamado *número de classe*. Se o número de classe é 1, a fatoração é única; caso contrário, não é. Quanto maior o número de classe, "menos única", em sentido significativo, é a fatoração em primos.

A capacidade de identificar a falta de singularidade da fatoração foi um grande passo adiante, e com um esforço extra conseguiu, às vezes, resgatar a estratégia de Lamé. Em 1850, Kummer anunciou que podia provar o último teorema de Fermat para uma grande quantidade de números primos, os que ele chamou de regulares. Entre os primos até 100, apenas 37, 59 e 67 são irregulares. Para todos os outros primos até esse limite, e muitos além dele, seus métodos provavam o último teorema de Fermat. A definição de números primos regulares requer o número de classe: um primo é regular se não divide o número de classe do correspondente anel de inteiros ciclotômicos. Assim, para um primo regular, embora sua fatoração não seja única, a maneira como ela falha em ser única não envolve o referido primo de forma essencial.

Kummer alegou que existem infinitos primos regulares, mas essa afirmativa permanece não provada. Ironicamente, em 1915, K.L. Jensen provou que existem infinitos primos irregulares. Um critério bizarro para um número primo ser regular emergiu das conexões com a análise, e envolve uma sequência de números descoberta de maneira independente pelo matemático japonês Seki Takazu (ou Kōwa) e pelo matemático suíço Jacob Bernoulli, chamada números de Bernoulli. O critério mostra que os dez primeiros primos irregulares são 37, 59, 67, 101, 103, 131, 149, 157, 233 e 257. Indo mais fundo na estrutura dos inteiros ciclotômicos, Dmitri Mirimanoff livrou-se do primeiro primo irregular, 37, em 1893. Em 1905, ele havia provado o último teorema de Fermat até $p = 257$. Harry Vandiver desenvolveu um algoritmo de computador que estendeu esse limite.

Usando tais métodos, John Selfridge e Bary Pollack provaram o teorema até a 25.000ª potência em 1967, e S. Wagstaff aumentou o teto para 100 mil em 1976.

A evidência da veracidade do último teorema de Fermat estava se acumulando, mas a principal implicação era que se o teorema fosse falso, então um contraexemplo – um exemplo exibindo que era falso – seria tão gigantesco que ninguém jamais seria capaz de encontrá-lo. A outra implicação era a de que métodos como o de Kummer acabam se deparando com os mesmos problemas que afligiram o trabalho dos pioneiros: potências maiores exigiam tratamento especial, mais complicado. Então essa linha de ataque foi lentamente perdendo força até cessar.

QUANDO VOCÊ ENCALHA em um problema matemático, o conselho de Poincaré é claro: deixe-o de lado e vá fazer outra coisa. Com sorte e bons ventos, uma nova ideia acabará por se revelar. Os teóricos dos números não seguiram o conselho de maneira consciente, mas mesmo assim fizeram o que ele havia recomendado. Como Poincaré insistiu que aconteceria, a tática deu certo. Alguns teóricos dos números voltaram sua atenção para as curvas elípticas (Capítulo 6). Ironicamente, essa área acabou revelando ter uma inesperada e surpreendente ligação com o último teorema de Fermat, levando à prova de Wiles. Para descrever esse elo, é necessário mais um conceito: o de função modular. A discussão vai ficar um pouco técnica, mas há uma história sábia por trás das ideias e precisaremos apenas de uma visão geral. Venha comigo.

No Capítulo 6, vimos que a teoria das funções elípticas teve um efeito profundo na análise complexa. Na década de 1830, Joseph Liouville descobriu que a variedade das funções elípticas é bastante limitada. Dados os dois períodos, há uma função elíptica especial, a função de Weierstrass, e toda outra função elíptica com esses dois períodos é uma simples variante. Isso significa que as únicas funções duplamente periódicas que se necessita entender são as funções de Weierstrass – uma para cada par de períodos.

Geometricamente, a estrutura duplamente periódica de uma função elíptica pode ser formulada em termos de uma grade no plano complexo: todas as combinações inteiras $mu + nv$ de dois períodos u e v, para inteiros m e n (Figura 30). Se pegarmos um número complexo z e somarmos a ele um desses pontos da grade, a função elíptica nesse novo ponto tem o mesmo valor que tinha no ponto original. Em outras palavras, a função

elíptica tem a mesma simetria que a grade.

Os analistas descobriram uma fonte muito mais rica de simetrias do plano complexo, conhecidas como transformações de Möbius. Estas mudam z para $(az + b)/(cz + d)$, para constantes complexas a, b, c, d. As simetrias na grade são tipos especiais de transformações de Möbius, mas há outras. Conjuntos de pontos análogos à grade ainda existem nessa disposição mais genérica. Uma grade define um padrão de ladrilhamento no plano euclidiano: use um paralelogramo como ladrilho e coloque seus vértices sobre os pontos da grade (Figuras 26 e 30). Usando as transformações de Möbius, podemos construir padrões de ladrilhamento numa geometria não euclidiana adequada, o plano hiperbólico. Podemos identificar essa geometria com uma região do plano complexo, no qual linhas retas são substituídas por arcos de círculos.

Existem padrões de ladrilhamento altamente simétricos na geometria hiperbólica. Para cada um deles, podemos construir funções complexas que repetem os mesmos valores em cada ladrilho. Estas são conhecidas como funções modulares, e geralmente são generalizações naturais de funções elípticas. A geometria hiperbólica é um tema muito rico, e a gama de padrões de ladrilhamento é muito mais extensa do que a do plano euclidiano. Assim, a análise complexa começou a pensar seriamente sobre geometria não euclidiana. Surgiu então um elo profundo entre análise e teoria dos números. As funções modulares representam para as curvas elípticas o que as funções trigonométricas representam para o círculo.

Lembremos que o círculo unitário consiste em pontos (x, y) tais que $x^2 + y^2 = 1$. Suponhamos que A seja um número real, e façamos

$$x = \cos A \qquad y = \operatorname{sen} A$$

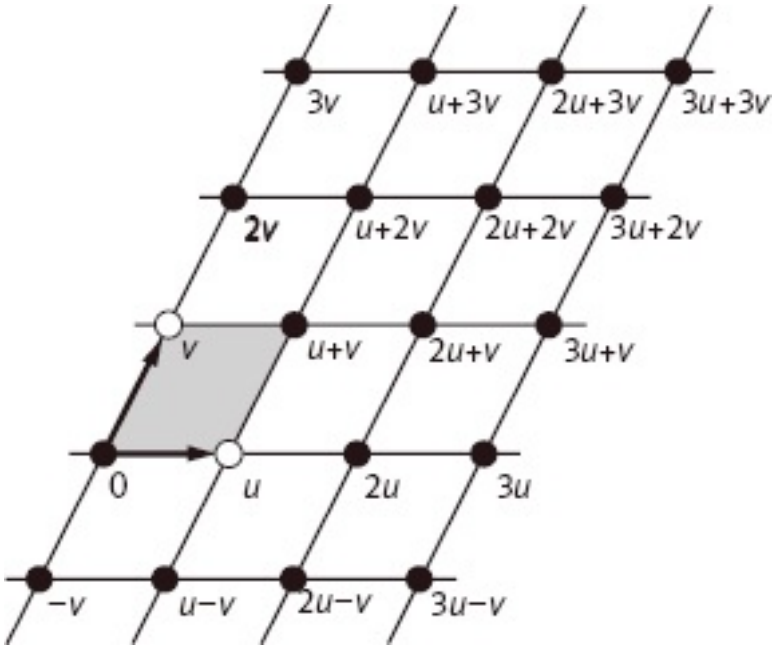


FIGURA 30 A grade é formada a partir de todas as combinações inteiras dos dois períodos.

Então a definição de seno e cosseno conta-nos que esse ponto está sobre um círculo unitário. Além disso, todo ponto no círculo unitário tem esse aspecto. No jargão, as funções trigonométricas *parametrizam* o círculo. Algo muito semelhante ocorre com as funções modulares. Se definirmos x e y usando funções modulares adequadas de um parâmetro A, o ponto correspondente fica sobre a curva elíptica, qualquer que seja o valor assumido por A. Há maneiras mais abstratas de tornar essa afirmação precisa, e os que trabalham na área as utilizam porque são mais convenientes, mas essa versão explicita a analogia com a trigonometria e o círculo. Essa conexão produz uma curva elíptica para cada função modular, e a variedade de funções modulares é imensa – todos ladrilhamentos simétricos do plano hiperbólico. Logo, uma quantidade enorme de curvas elípticas pode ser relacionada com funções modulares. Que curvas elípticas podem ser obtidas desse modo? Este acabou se revelando o cerne da questão.

ESTE "ELO PERDIDO" ganhou proeminência pela primeira vez em 1975, quando Yves Hellegouarch notou uma curiosa conexão entre o último teorema de Fermat e as curvas elípticas. Gerhard Frey desenvolveu a ideia em dois artigos publicados em 1982 e 1986. Suponhamos, como sempre, que p é um primo ímpar. Vamos admitir – na esperança de deduzir uma contradição – que existam inteiros a, b e c diferentes de zero que satisfaçam a equação de

Fermat, de modo que $a^p + b^p = c^p$. Agora vamos tirar o coelho da cartola com um floreio teatral: consideremos a curva elíptica

$$y^2 = x(x - a^p)(x - b^p)$$

Esta é chamada de curva elíptica de Frey. Frey aplicou o mecanismo das curvas elípticas a ela, e o que surgiu foi uma cadeia de coincidências ainda mais bizarras. Sua curva elíptica hipotética é de fato muito estranha. Ela parece não fazer sentido. Frey provou que ela faz tão pouco sentido que não pode existir. E isso, é claro, prova o último teorema de Fermat, fornecendo a requerida contradição.

No entanto, havia uma lacuna, da qual Frey estava bem ciente. Para provar que a sua curva hipotética não existe, é preciso demonstrar que, se existisse, seria modular – ou seja, uma das curvas que surgem a partir das funções modulares. Acabamos de ver que tais curvas são comuns, e naquela época ninguém jamais havia encontrado uma curva elíptica que *não* fosse modular. Parecia provável que a curva de Frey devesse ser modular – mas era uma curva hipotética, os números a, b, c eram desconhecidos, e se a curva *fosse* modular, então ela não existiria de jeito nenhum. Contudo, havia um modo de lidar com todas essas questões: provar que *toda* curva elíptica é modular. Então a curva de Frey, hipotética ou não, teria de ser modular se existisse. E se não existisse, de qualquer modo a prova estava completa.

A afirmação de que toda curva elíptica é modular chama-se conjectura de Taniyama-Shimura. Ela recebe o nome de dois matemáticos japoneses, Yutaka Taniyama e Goro Shimura. Eles se encontraram por acaso, ambos querendo tirar o mesmo livro na biblioteca ao mesmo tempo pela mesma razão. Isso deflagrou uma longa colaboração. Em 1955, Taniyama encontrava-se em uma conferência de matemática em Tóquio e os jovens participantes foram convidados a reunir uma lista de questões em aberto. Taniyama contribuiu com quatro, todas apontando para uma relação entre funções modulares e curvas elípticas. Ele havia calculado alguns números associados a uma função modular específica, e notado que exatamente os mesmos quatro números apareciam em conexão com uma particular curva elíptica. Esse tipo de coincidência é muitas vezes um sinal de que não se trata efetivamente de uma coincidência, de que deve haver alguma explicação razoável. A constância desses números, sabe-se agora, é equivalente ao fato de a curva elíptica ser modular; na verdade, essa é a definição preferida na literatura de pesquisa. Em todo caso, Taniyama ficou

suficientemente intrigado por ter calculado os números para algumas outras funções modulares, descobrindo que também estes correspondiam a curvas elípticas específicas.

Ele começou a se perguntar se algo similar funcionaria para toda curva elíptica. A maioria dos atuantes no campo considerou isso bom demais para ser verdade, um castelo no ar para o qual havia muito pouca evidência. Shimura foi um dos poucos que sentiram que a conjectura tinha mérito. Mas em 1957-58, Shimura passou um ano em Princeton e, enquanto estava fora, Taniyama cometeu suicídio. Ele deixou um bilhete que em parte dizia: "Quanto à causa do meu suicídio, eu mesmo não a entendo muito bem, mas não é resultado de um acidente particular nem de um assunto específico. Meramente posso dizer, estou num estado de espírito em que perdi a confiança no meu futuro." Na época ele estava planejando se casar, e a futura noiva, Misako Suzuki, matou-se um mês depois. Seu bilhete de suicídio incluía: "Agora que ele se foi, preciso ir também para juntar-me a ele."

Shimura continuou a trabalhar na conjectura, e à medida que se acumulava evidência a seu favor, começou a pensar que poderia realmente ser verdade. A maioria dos outros que trabalhavam na área discordava. Simon Singh[4] relata uma entrevista com Shimura, na qual ele se recordava de tentar explicar isso a um de seus colegas:

O professor inquiriu: "Ouço dizer que o senhor propôs que algumas equações elípticas podem ser ligadas a formas modulares."

"Não, não, o senhor não entende", replicou Shimura. "Não são apenas *algumas* equações elípticas, é *toda* equação elíptica!"

A despeito desse tipo de ceticismo, Shimura perseverou, e após muitos anos a proposta tornou-se suficientemente respeitável para ser citada como conjectura Taniyama-Shimura. Então André Weil, um dos grandes teóricos de números do século XX, encontrou muito mais evidências em favor da conjectura, divulgou-a e manifestou a crença de que poderia muito bem ser verdadeira. Ela tornou-se conhecida como conjectura Taniyama-Shimura-Weil. O nome nunca pegou direito, e muitas permutações de subconjuntos dos três matemáticos foram associadas a ela. Vamos ficar com "conjectura Taniyama-Shimura".

Na década de 1960, outro peso-pesado, Robert Langlands, percebeu que

a conjectura Taniyama-Shimura podia ser vista como apenas mais um elemento num programa mais amplo e mais ambicioso, que unificaria a teoria dos números algébrica e analítica. Ele semeou toda uma plantação de conjecturas relacionadas com essa ideia, agora conhecida como programa Langlands. Era algo ainda mais especulativo que a conjectura Taniyama-Shimura, mas tinha uma fascinante elegância, o tipo de matemática que precisava ser verdadeira porque era bela. Ao longo dos anos 1970, o mundo da matemática acostumou-se à beleza selvagem do programa Langlands, que passou a ser aceito como um dos alvos centrais da teoria algébrica dos números. O programa Langlands parecia o caminho correto para seguir em frente, só faltava alguém dar o primeiro passo.

Nesse ponto, Frey notou que aplicar a conjectura de Taniyama-Shimura à sua curva elíptica provaria o último teorema de Fermat. No entanto, a essa altura viera à tona outro problema com a ideia de Frey. Quando proferiu uma palestra sobre o tema em 1984, a plateia identificou uma lacuna em seu argumento-chave: a curva é tão bizarra que não pode ser modular. Jean-Pierre Serre, uma das principais figuras da área, rapidamente preencheu a lacuna, mas precisou invocar outro resultado que também carecia de prova, a conjectura da redução de nível especial. Em 1986, porém, Ken Ribet havia provado a conjectura da redução de nível especial. Agora o único obstáculo para a prova do último teorema de Fermat era a conjectura de Taniyama-Shimura, e o consenso começou a mudar. Serre predisse que o último teorema de Fermat provavelmente seria provado no período de uma década ou algo assim. Exatamente como era outra questão, mas havia uma sensação generalizada no ar: as técnicas relacionadas com funções modulares estavam se tornando tão poderosas que alguém em breve faria a abordagem de Frey funcionar.

ESSE ALGUÉM FOI Andrew Wiles. No programa de televisão sobre sua prova, ele disse:

Eu tinha dez anos e … encontrei um livro de matemática que contava um pouco da história desse problema [o último teorema de Fermat] – que alguém havia [apresentado] este problema trezentos anos atrás, mas ninguém jamais tinha visto uma prova, ninguém sabia se havia prova, e desde então as pessoas andavam à procura da prova. E aqui estava um problema que eu, um menino de dez anos, podia entender, mas nenhum

dos grandes matemáticos do passado fora capaz de resolver. E a partir desse momento, é claro, simplesmente tentei resolvê-lo eu mesmo. Era um desafio tão grande, um problema tão bonito.

Em 1971, Wiles graduou-se em matemática em Oxford e mudou-se para Cambridge, para o doutorado. Seu supervisor, John Coates, o advertiu (corretamente) que o último teorema de Fermat era difícil demais para um doutorado. Então, em vez disso, Wiles se pôs a trabalhar com curvas elípticas, na época considerada uma área de pesquisa muito mais promissora. Em 1985-86, ele estava em Paris no Institut des Hautes Études Scientifiques, um dos institutos de pesquisa mais avançados do mundo. A maioria dos pesquisadores de ponta passou por ele em algum momento. Se você é matemático, é um ótimo lugar para badalar. Entre os visitantes estava Ribet, e sua prova da conjectura de redução de nível especial deixou Wiles eletrizado. Agora ele podia mergulhar numa pesquisa inteiramente respeitável das curvas elípticas, tentando provar a conjectura de Taniyama-Shimura e ao mesmo tempo tentar realizar seu sonho de infância provando o último teorema de Fermat.

Como todos da área agora sabiam da conexão, havia uma preocupação. Vamos supor que Wiles conseguisse montar uma prova quase completa, com pequenas lacunas exigindo um pouco de trabalho adicional. Vamos supor que alguém ficasse sabendo disso e preenchesse as lacunas. Então, tecnicamente, a prova do último teorema de Fermat seria dessa pessoa. Os matemáticos geralmente procuram não comportar-se dessa maneira, mas quando o prêmio é grande, é sábio tomar precauções. Assim, Wiles realizou sua pesquisa em segredo, algo que os matemáticos raramente fazem. Não é que não confiasse nos colegas. Apenas não podia correr o mínimo risco de perder a corrida por um fio de cabelo.

Trabalhou por sete anos enfurnado no sótão da sua casa, onde havia um escritório. Somente sua esposa e seu chefe de departamento sabiam no que ele estava trabalhando. Em paz e em reclusão, atacou o problema com toda técnica que pudesse descobrir, até que os muros do castelo começaram a tremer sob o bombardeio. Em 1991, Coates o colocou a par de alguns novos resultados provados por Mattheus Flach. A rachadura no muro começou a se alargar mais depressa à medida que o assédio se intensificava.

Em 1993, a prova estava completa. Agora precisava ser revelada para o mundo. Ainda cauteloso, Wiles não quis arriscar proclamar a solução apenas para algum erro vir à tona – algo que ocorrera com Yoichi Miyaoka

em 1988, cuja reivindicação da prova chegou à mídia, apenas para um erro fatal ser encontrado. Assim, Wiles decidiu dar uma série de três palestras no Isaac Newton Institute em Cambridge, um recém-fundado centro internacional de pesquisa matemática. O título era inócuo e técnico: "Formas modulares, curvas elípticas e teoria de Galois." O que conseguiu enganar bem pouca gente: sabia-se que Wiles estava para revelar algo grandioso.

Na terceira palestra, Wiles esboçou uma prova de um caso especial da conjectura de Taniyama-Shimura. Ele descobrira que algo um pouco menos ambicioso também funcionaria. Provar que a curva de Frey, caso exista, deve pertencer a uma classe especial de curvas elípticas, as "semiestáveis", e provar que todas as curvas *nessa classe* devem ser modulares. Wiles então provou ambos os resultados. No fim da palestra, escreveu um corolário – um teorema suplementar que se segue diretamente ao que quer que tenha sido provado – no quadro-negro. O corolário era o último teorema de Fermat.

Quando Shimura ouviu falar do anúncio de Wiles, seu comentário foi breve e direto ao ponto: "Eu bem que disse."

SE AO MENOS TIVESSE sido tão simples e direto. Mas o destino ainda tinha guardado um pequeno tropeço. A prova ainda precisava ser referendada por peritos, e, como de hábito, o processo revelou alguns pontos que necessitavam de explicação adicional. Wiles lidou com a maioria dos comentários, mas um deles o forçou a repensar. No fim de 1993, ele emitiu um comunicado dizendo que estava retirando sua reivindicação até costurar uma lacuna lógica que havia surgido. Mas agora era forçado a operar sob as luzes plenas da publicidade, exatamente o que tinha esperado evitar.

Em março de 1994, nenhuma prova consertada aparecera, e Faltings expressou uma opinião difundida na comunidade matemática: "Se [consertar a prova] fosse fácil, a esta altura ele já a teria resolvido. Estritamente falando, não foi uma prova ao ser anunciada." Weil comentou: "Acredito que ele tenha algumas boas ideias … mas a prova não está lá … provar o último teorema de Fermat é como escalar o Everest. Se um homem quer escalar o Everest e desiste quando faltam cem metros, ele não escalou o Everest." Todo mundo já podia adivinhar como a coisa acabaria. Já tinham visto antes. A prova caíra por terra, teria de ser completamente

recolhida, e o último teorema de Fermat viveria mais algum tempo.

Wiles recusou-se a reconhecer a derrota, e seu ex-aluno Richard Taylor juntou-se a ele na busca. A raiz da dificuldade agora estava clara: os resultados de Flach não eram exatamente adequados para a tarefa. Ambos tentaram modificar os métodos de Flach, mas nada parecia dar certo. Então, num rasgo de inspiração, Wiles de repente compreendeu qual era o obstáculo. "Eu vi que o que tinha impedido [o método de Flach] de dar certo era algo que faria funcionar outro método que eu havia tentado anteriormente." Foi como se soldados sitiando o castelo tivessem subitamente percebido que seu aríete jamais funcionaria porque os defensores ficavam jogando pedras sobre ele, mas essas mesmíssimas pedras poderiam abastecer uma catapulta e ser usadas para derrubar o portão.

Em abril de 1995, a nova prova estava terminada, e dessa vez não havia lacunas nem erros. A publicação veio rapidamente: dois artigos na ultraprestigiosa *Annals of Mathematics*. Wiles tornou-se uma celebridade internacional, tendo recebido diversos prêmios importantes além de ser sagrado cavaleiro… E voltou à sua pesquisa, comportando-se de forma muito parecida com a anterior.

A CARACTERÍSTICA REALMENTE importante da solução de Wiles não é absolutamente o último teorema de Fermat. Como eu já disse, nada de vital advém da resposta. Se alguém tivesse encontrado três números de cem dígitos e um primo de 250 dígitos que fornecessem um contraexemplo da alegação de Fermat, então o teorema seria falso, mas nada de crucial na matemática ficaria de forma alguma diminuído. É claro que um ataque direto por computador não conseguiria vasculhar números tão grandes, então você teria de ser assombrosamente sagaz para estabelecer tal coisa, mas um resultado negativo não provocaria nenhum ataque cardíaco.

A importância real da solução reside na prova do caso semiestável da conjectura de Taniyama-Shimura. Durante seis anos, Christophe Breuil, Brian Conrad, Fred Diamond e Taylor estenderam os métodos de Wiles para lidar não só com o caso semiestável, mas com todas as curvas elípticas. Eles provaram integralmente a conjectura Taniyama-Shimura, e a teoria dos números jamais seria a mesma. Desse ponto em diante, sempre que alguém encontrasse uma curva elíptica, ela tinha a garantia de ser modular, então viria se abrir uma horda de métodos analíticos. Esses

métodos têm sido usados para resolver outros problemas em teoria dos números e mais ainda aparecerão no futuro.

8. Caos orbital

O problema dos três corpos

SEGUNDO UMA ANTIGA e consagrada piada, pode-se saber o grau de adiantamento de uma teoria física pelo número de corpos interagindo com os quais ela consegue lidar. A lei da gravitação de Newton enfrenta problemas com três corpos. A relatividade geral tem dificuldade em tratar com dois corpos. A teoria quântica já é superestendida para um corpo, e a teoria quântica de campo mete-se em apuros com *nenhum* corpo – o vácuo. Como muitas piadas, esta contém um grão de verdade.[1] Em particular, a interação gravitacional de meros três corpos, assumida como obedecendo à lei da gravitação do inverso do quadrado, formulada por Newton, desafiou o mundo matemático durante séculos. E ainda desafia, se você quiser uma fórmula bacana para as órbitas desses corpos. Na verdade, agora sabemos que a dinâmica de três corpos é caótica – tão irregular que tem elementos de aleatoriedade.

Tudo isso está em enorme contraste com o assombroso sucesso da teoria gravitacional de Newton, que explicou, entre muitas outras coisas, a órbita de um planeta ao redor do Sol. A resposta é aquilo que Kepler já deduzira empiricamente a partir de observações astronômicas de Marte: uma elipse. Aqui ocorrem apenas dois corpos: o Sol e o planeta. O passo seguinte óbvio é usar a lei da gravitação de Newton para anotar a equação para órbitas de três corpos, e resolvê-la. Mas não existe caracterização geométrica organizada de órbitas de três corpos, nem mesmo uma fórmula em geometria de coordenadas. Até o fim do século XIX, muito pouco se sabia a respeito do movimento de três corpos celestes, mesmo que um deles fosse tão pequeno que sua massa pudesse ser ignorada.

Nossa compreensão da dinâmica de três (ou mais) corpos tem crescido drasticamente desde então. Grande parte desse progresso tem sido uma crescente percepção de quanto essa questão é difícil, e por quê. Pode parecer um retrocesso, mas às vezes a melhor maneira de avançar é fazer um recuo estratégico e tentar algo diferente. Para o problema dos três corpos, este plano de campanha tem trazido alguns sucessos reais, quando um ataque de frente teria sido irremediavelmente rechaçado.

OS PRIMEIROS SERES HUMANOS não puderam deixar de notar que a Lua se move gradualmente através do céu noturno, em relação ao fundo de estrelas. As estrelas também parecem mover-se, mas o fazem como um bloco inteiro, como minúsculas cabeças de alfinete feitas de luz sobre uma vasta tigela giratória. A Lua é claramente especial sob outro aspecto: é um grande disco brilhante, que muda de forma, da lua nova para a lua cheia e de volta para a nova. Não é um pontinho de luz como uma estrela.

Alguns desses pontinhos de luz também desobedecem às regras. Eles se movem desordenadamente. Não mudam de posição em relação ao fundo de estrelas com a mesma rapidez que a Lua, mas mesmo assim não é preciso observar o céu por muitas noites para ver que alguns estão se movendo. Cinco desses pontinhos nômades são visíveis a olho nu; os gregos os chamaram de *planetes* – nômades. São, obviamente, os planetas, e os cinco que foram reconhecidos desde os tempos antigos são os que agora chamamos de Mercúrio, Vênus, Marte, Júpiter e Saturno – todos com nomes de deuses romanos. Com auxílio de telescópios sabemos agora de mais dois: Urano e Netuno. Mais a nossa Terra, é claro. Plutão não conta mais como planeta, graças a uma controvertida decisão sobre terminologia tomada pela União Astronômica Internacional em 2006.

À medida que os antigos filósofos, astrônomos e matemáticos foram estudando os planetas, perceberam que eles não se movem simplesmente ao acaso. Seguem trajetos retorcidos, mas bastante previsíveis, e retornam praticamente à mesma posição no céu noturno em intervalos de tempo bastante regulares. Agora explicamos esses padrões como o movimento periódico percorrendo uma órbita fechada, com uma pequena contribuição do próprio movimento orbital da Terra. Reconhecemos também que a periodicidade não é exata – mas chega perto. Mercúrio leva aproximadamente 88 dias para dar a volta ao redor do Sol, enquanto Júpiter leva quase doze anos. Quanto mais longe do Sol o planeta está, mais tempo gasta para completar sua órbita.

O primeiro modelo quantitativamente acurado do movimento dos planetas foi o sistema ptolemaico, assim batizado em razão de Claudio Ptolomeu, que o descreveu em seu *Almagesto* (O maior [tratado]), de cerca do ano 150. É um modelo geocêntrico – com centro na Terra –, no qual todos os corpos celestes orbitam em volta da Terra. Movem-se como que sustentados por uma série de esferas gigantes, cada um girando num ritmo fixo em torno de um eixo que pode ser, ele próprio, sustentado por outra esfera. Eram necessárias combinações de muitas esferas giratórias para

representar o complexo movimento dos planetas em termos do ideal cósmico de rotação uniforme num círculo – o equador da esfera. Com esferas suficientes e a escolha correta de seus eixos e velocidades, o modelo corresponde de perto à realidade.

Nicolau Copérnico modificou o esquema de Ptolomeu de diversas maneiras. A mais radical foi fazer todos os corpos, com exceção da Lua, girar em torno do Sol, o que simplificou consideravelmente a descrição. Era um modelo heliocêntrico. Essa proposta foi mal recebida pela Igreja católica, mas o ponto de vista científico acabou prevalecendo, e as pessoas cultas aceitaram que a Terra gira em torno do Sol. Em 1596, Kepler defendeu o sistema de Copérnico em seu *Mysterium Cosmographicum* (O mistério cosmográfico), cujo ponto alto foi a descoberta de uma relação matemática entre a distância do planeta ao Sol e seu período orbital. Movendo-se para longe do Sol, a razão de aumento no período de um planeta para o planeta seguinte é o dobro do aumento da distância. Mais tarde, ele concluiu que essa relação era imprecisa demais para estar correta, mas lançou as sementes de uma relação mais acurada em seu futuro trabalho. Kepler também explicou o espaçamento dos planetas em termos dos cinco sólidos regulares, aninhados precisamente um dentro do outro, separados pelas esferas que os sustentavam. Cinco sólidos explicavam por que havia cinco planetas, mas agora reconhecemos oito, então essa característica não é mais vantagem. Há 120 maneiras diferentes de organizar cinco sólidos, e uma delas tem a probabilidade de chegar perto das proporções celestes dadas pelas órbitas planetárias. Então a aproximação é apenas acidental, ajustando forçadamente a natureza num padrão sem significado.

Em 1600, o astrônomo Tycho Brahe contratou Kepler para ajudá-lo a analisar suas observações, mas houve a intervenção de problemas políticos. Após a morte de Brahe, Kepler foi nomeado matemático imperial de Rodolfo II. Em seu tempo livre, trabalhava nas observações de Marte feitas por Brahe. Um dos resultados foi *Astronomia Nova*, de 1609, que apresentou mais duas leis do movimento planetário. A primeira lei de Kepler afirma que os planetas movem-se em elipses – ele estabelecera isso para Marte, e parecia provável que o mesmo seria verdade para os outros planetas. A princípio presumiu que um formato de ovo se encaixaria nos dados, mas não funcionou, então tentou a elipse. Esta também foi rejeitada e ele encontrou uma descrição matemática diferente do formato da órbita. Finalmente descobriu que era na verdade outro modo de definir uma

elipse:[2]

Deixei [a nova definição] de lado e caí de novo nas elipses, acreditando que esta era uma hipótese bem diferente, quando as duas, como provarei no próximo capítulo, são uma só na mesma … Ah, que pássaro tolo tenho sido!

A segunda lei de Kepler afirma que o planeta varre áreas iguais em tempos iguais. Em 1619, em seu *Harmonices Mundi* (Harmonias do mundo), Kepler completou suas três leis com uma relação bem mais precisa entre distâncias e períodos: o cubo da distância (metade do comprimento do eixo maior da elipse) é proporcional ao quadrado do período.

O palco estava agora montado para Isaac Newton. Em seu *Philosophiae Naturalis Principia Mathematica* (Princípios matemáticos da filosofia natural), de 1687, Newton provou que as três leis de Kepler são equivalentes a uma única lei de gravitação: dois corpos atraem-se mutuamente com uma força que é proporcional a suas massas e inversamente proporcional ao quadrado da distância entre eles. A lei de Newton tinha uma enorme vantagem: aplicava-se a qualquer sistema de corpos, por maior que fosse seu número. O preço a ser pago era a maneira como a lei prescrevia as órbitas: não como formas geométricas, mas como soluções de uma equação diferencial, que envolvia as acelerações dos planetas. Não fica totalmente claro como encontrar, a partir dessa equação, os formatos das órbitas planetárias, ou as posições dos planetas num dado momento. Em poucas palavras, não fica absolutamente claro como achar suas acelerações. No entanto, a equação fornecia *implicitamente* essa informação. O problema era deixá-la explícita. Kepler já o fizera para dois corpos, e a resposta era órbitas elípticas percorridas com velocidades que varriam áreas numa razão constante.

E quanto a três corpos?

ERA UMA BOA PERGUNTA. Segundo a lei de Newton, todos os corpos no sistema solar influenciam-se gravitacionalmente uns aos outros. Na verdade, todos os corpos no universo inteiro influenciam-se gravitacionalmente uns aos outros. Mas ninguém em sã consciência tentaria escrever equações diferenciais para cada corpo no universo. Como sempre, o caminho a seguir era simplificar o problema – mas não demais. As estrelas estão tão distantes que seu efeito gravitacional sobre o sistema solar é desprezível, a

menos que se queira descrever como o Sol se move à medida que a galáxia gira. O movimento da Lua é basicamente influenciado por dois outros corpos: a Terra e o Sol, com exceção de alguns efeitos sutis envolvendo outros planetas. No começo do século XVIII, essa questão fugiu do reino da astronomia e adquiriu importância prática, quando se percebeu que o movimento da Lua podia ser útil para a navegação. (Não havia GPS naquele tempo, nem mesmo instrumentos para medir a longitude.) Mas esse método requeria predições mais acuradas do que as teorias existentes eram capazes de prover. O lugar óbvio para se começar era escrever as implicações da lei de Newton para três corpos, os quais, para este propósito, podiam ser tratados como massas puntiformes, pois os planetas são extremamente pequenos em comparação com as distâncias entre eles. Então resolver-se-iam as equações diferenciais. No entanto, os truques que levavam de dois corpos para as elipses falhavam quando um terceiro corpo entrava na composição. Alguns passos preliminares funcionavam, mas então o cálculo se deparava com uma obstrução. Em 1747, Jean d'Alembert e Alexis Clairaut, amargos rivais, competiram por um prêmio da Academia de Ciências de Paris sobre o "*problème des trois corps*", que ambos abordaram por meio de aproximações numéricas. O problema dos três corpos tinha adquirido seu nome, e logo se tornou um dos maiores enigmas em matemática.

Alguns casos especiais puderam ser resolvidos. Em 1767, Euler descobriu soluções nas quais todos os três corpos estão dispostos numa linha reta giratória. Em 1772, Lagrange encontrou soluções similares em que os três corpos formam um triângulo equilátero giratório, que se expande e se contrai. Ambas as soluções eram periódicas: os corpos repetiam indefinidamente a mesma sequência de movimentos. Porém, mesmo simplificações drásticas fracassaram em produzir algo mais genérico. Podia-se assumir que um dos corpos tivesse massa desprezível, podia-se assumir que os outros dois moviam-se em círculos perfeitos em torno de seu centro de massa mútuo, versão conhecida como problema dos três corpos "restrito"… E *ainda assim* não se podiam resolver as equações com exatidão.

Em 1860 e 1867, o astrônomo e matemático Charles-Eugène Delaunay atacou o caso específico do sistema Sol-Terra-Lua usando a teoria das perturbações, que encara o efeito da gravidade do Sol sobre a Lua como uma pequena mudança imposta sobre o efeito da Terra, deduzindo fórmulas aproximadas em forma de série: a soma de muitos termos

sucessivos. Delaunay publicou seus resultados em 1860 e 1867; cada volume tinha novecentas páginas, consistindo em grande parte de fórmulas. No fim dos anos 1970, seus cálculos foram conferidos usando-se álgebra computacional, encontrando-se apenas dois erros pequenos e sem importância.

Foi um cálculo heroico, mas a série tendia ao seu valor limite muito lentamente para ter alguma utilidade prática. Todavia, incentivou outros a buscar soluções de séries que convergissem mais rapidamente. E também revelou um grande obstáculo técnico para todas essas abordagens, conhecido como pequenos denominadores. Alguns termos da série são frações, e o denominador (a parte de baixo da fração) torna-se muito pequeno se os corpos estão próximos de uma ressonância: um estado periódico no qual os períodos são múltiplos racionais um do outro. Por exemplo, as três luas mais internas de Júpiter, Io, Europa e Ganimedes, têm períodos de rotação em volta do planeta de 1,77 dia, 3,55 dias e 7,15 dias, numa razão quase exata de 1 : 2 : 4. Ressonâncias seculares, relações racionais entre as razões em que giram os eixos de duas órbitas quase elípticas, são um aborrecimento especial, porque o erro provável em avaliar uma fração torna-se muito grande quando o denominador é pequeno.

SE O PROBLEMA dos três corpos era difícil, o problema de *n* corpos – qualquer quantidade de massas puntiformes movendo-se por gravitação newtoniana – seguramente era mais difícil ainda. Todavia, a natureza nos apresenta um importante exemplo: o sistema solar inteiro. Este contém oito planetas, vários planetas anões como Plutão e milhares de asteroides, alguns bastante grandes. Para não mencionar os satélites, alguns dos quais – Titã, por exemplo – são maiores que o planeta Mercúrio. Logo, o sistema solar é um problema de dez corpos, ou um problema de vinte corpos, ou um problema de mil corpos, dependendo de quanto detalhe você queira incluir.

Para previsões de curto prazo, aproximações numéricas são eficazes, e, em astronomia, mil anos é curto prazo. Compreender como o sistema solar vai evoluir no passar de centenas de milhares de anos é algo bem diferente. E uma grande questão depende desse tipo de visão de longo prazo: a estabilidade do sistema solar. Os planetas parecem estar se movendo em órbitas relativamente estáveis, quase elípticas. As órbitas mudam um pouquinho quando outros planetas as perturbam, então o período pode

mudar uma fração de segundo, ou o tamanho da elipse pode não ser exatamente constante. Podemos ter certeza de que esses suaves cutucões são tudo que vai ocorrer no futuro? É típico do que ocorreu no passado, sobretudo nos estágios iniciais do sistema solar? Será que o sistema solar vai permanecer estável, ou dois planetas colidirão? Será que um planeta poderia ser lançado para os distantes confins do universo?

O ano de 1889 foi o sexagésimo aniversário de Oscar II, rei da Noruega e da Suécia. Como parte das celebrações, o matemático norueguês Gösta Mittag-Leffler persuadiu o monarca a anunciar um prêmio para a solução do problema de *n* corpos. Esta deveria ser conseguida não por uma fórmula exata – a essa altura já estava claro que isso seria pedir demais –, mas por algum tipo de série convergente. Poincaré interessou-se, e resolveu começar com uma versão bem simples: o problema restrito dos três corpos, onde um tem massa desprezível, como uma ínfima partícula de poeira. Se você aplicar de maneira ingênua a lei de Newton a essa partícula, a força exercida sobre ela é o produto das massas dividido pelo quadrado da distância, e uma das massas é zero, então o produto é zero. Isso não adianta muito, porque a partícula de poeira simplesmente segue seu próprio caminho, desligada dos dois outros corpos. Em vez disso, ajusta-se o modelo de maneira tal que a partícula sinta o efeito dos outros dois corpos, mas estes a ignoram completamente. Assim, as órbitas dos dois corpos maciços são circulares, e eles se movimentam numa velocidade fixa. Toda a complexidade do movimento é investida na partícula de poeira.

Poincaré não solucionou o problema apresentado pelo rei Oscar. Esse era ambicioso demais. Mas seus métodos foram tão inovadores, e fizeram tanto progresso, que de todo modo o prêmio lhe foi conferido. Sua pesquisa premiada foi publicada em 1890, e sugeria que mesmo o problema restrito dos três corpos poderia não possuir o tipo de resposta que fora estipulada. Poincaré dividiu sua análise em vários casos distintos, dependendo das características gerais do movimento. Na maioria deles, soluções de séries podiam muito bem ser obtidas. Mas havia um caso no qual a órbita da partícula de poeira ficava extremamente confusa.

Poincaré deduziu essa confusão inevitável de algumas outras ideias que vinha desenvolvendo, que possibilitavam descrever soluções para equações diferenciais sem efetivamente resolvê-las. Essa “teoria qualitativa de equações diferenciais” foi a semente da qual se desenvolveu a moderna dinâmica não linear. A ideia básica era explorar a geometria das soluções; mais precisamente, sua topologia, um tópico no qual Poincaré também

estava profundamente interessado (Capítulo 10). Nessa interpretação, as posições e velocidades dos corpos são coordenadas num espaço multidimensional. À medida que o tempo passa, qualquer estado inicial segue uma trajetória curva através desse espaço. A topologia dessa trajetória, ou o sistema inteiro de todas as trajetórias possíveis, nos conta muitas coisas úteis sobre as soluções.

Uma solução periódica, por exemplo, é uma trajetória que se fecha sobre si mesma de modo a formar um laço. Com o passar do tempo, o estado dá voltas e voltas pelo laço, repetindo indefinidamente o mesmo comportamento. O sistema é então periódico. Poincaré sugeriu que uma boa maneira de detectar tais laços é dispor uma superfície multidimensional de modo que ela faça um corte através do laço. Nós agora a chamamos de seção de Poincaré. Soluções que começam nessa superfície podem eventualmente retornar a ela, e soluções através de pontos vizinhos sempre retornam à seção após aproximadamente um período. Logo, uma solução periódica pode ser interpretada como um ponto fixo do "primeiro mapa de retorno", que nos diz o que acontece com pontos na superfície quando voltam a ela pela primeira vez, se isso ocorrer. Pode não parecer um grande avanço, mas reduz a dimensão do espaço – o número de variáveis do problema. Isso quase sempre é algo bom.

A grande ideia de Poincaré começa a tomar forma quando passamos para o seguinte tipo de solução em termos de complexidade, combinações de diversos movimentos periódicos. Como exemplo simples, a Terra gira em torno do Sol aproximadamente a cada 365 dias e a Lua gira em torno da Terra aproximadamente a cada 28 dias. Logo, o movimento da Lua combina esses dois períodos diferentes. É claro que toda a questão do problema dos três corpos é que essa descrição não é inteiramente precisa, mas soluções "quase periódicas" desse tipo são bastante comuns nos problemas de muitos corpos. A seção de Poincaré detecta soluções quase periódicas: quando retornam à superfície elas não atingem exatamente o mesmo ponto, mas o ponto que atingem move-se ao longo de uma curva fechada *sobre a superfície*, em pequenos passos.

Poincaré percebeu que se toda solução fosse como essa, ele seria capaz de montar séries adequadas para modelar tais soluções quantitativamente. Mas quando analisou a topologia dos primeiros mapas de retorno, notou que podia ser mais complicado. Duas curvas específicas, relacionadas pela dinâmica, poderiam se cruzar. Isso em si não era algo tão ruim, mas quando eram espichadas até voltarem a atingir a superfície, as curvas resultantes

ainda tinham de se cruzar – mas num local diferente. Force mais uma volta, e elas se cruzam de novo. E não só isso: essas novas curvas que surgiram espichando-se as curvas originais não eram realmente novas. Eram partes das curvas originais. Descobrir a topologia exigia pensar com clareza, porque ninguém realmente havia feito esse tipo de jogo antes. O que surge é uma figura muito complexa, como uma rede maluca, na qual as curvas ziguezagueiam repetidamente de um lado a outro, cruzando-se, e os próprios zigue-zagues ziguezagueiam de um lado a outro, e assim por diante, em qualquer nível de complexidade. Poincaré, com efeito, declarou-se perplexo:

Quando se tenta descrever a figura formada por estas duas curvas e sua infinidade de intersecções, cada uma correspondendo a uma solução duplamente assintótica, essas intersecções formam uma espécie de rede, trama ou malha infinitamente apertada … Fica-se impressionado com a complexidade dessa figura que eu nem mesmo tento desenhar.

Chamaremos então essa figura de emaranhado homoclínico ("autoconectado") (Figura 31). Graças a novas ideias topológicas introduzidas nos anos 1960 por Stephen Smale, reconhecemos agora essa estrutura como uma velha amiga. Sua implicação mais importante é que a dinâmica é *caótica*. Embora as equações não tenham nenhum elemento explícito de aleatoriedade, suas soluções são muito complicadas e irregulares, compartilhando certas características de processos genuinamente aleatórios. Por exemplo, há órbitas – na verdade, a maioria delas – para as quais o movimento imita exatamente o lançar aleatório de uma moeda. A descoberta de que um sistema determinístico – um sistema cujo futuro inteiro é exclusivamente determinado por seu estado presente – pode ter, não obstante, características aleatórias é algo extraordinário, e provocou mudanças em muitas áreas da ciência. Não assumimos mais automaticamente que regras simples causem comportamento simples. Trata-se do que é conhecido de modo coloquial como teoria do caos, e tudo remonta a Poincaré e seu prêmio Oscar.

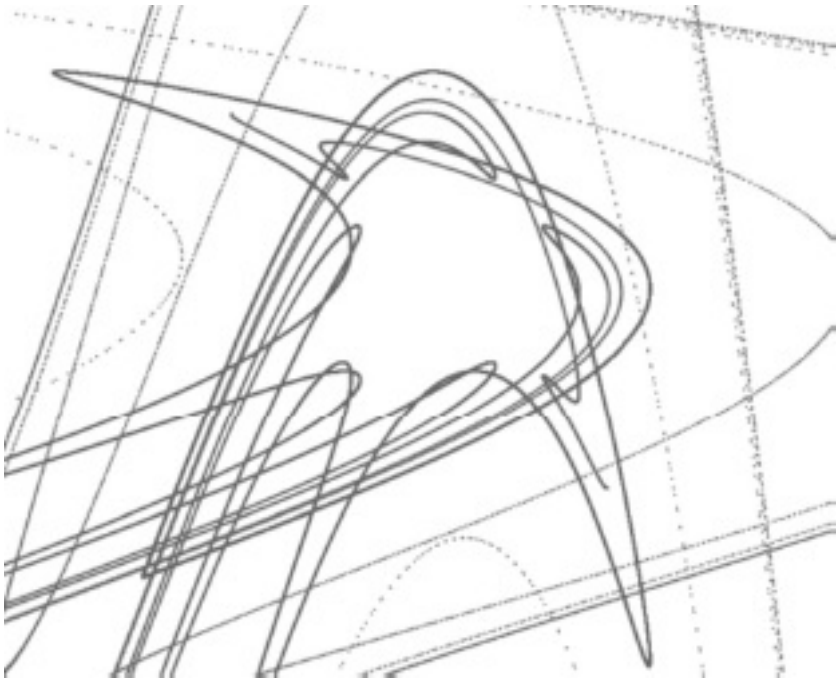


FIGURA 31 Parte de um emaranhado homoclínico. Uma figura completa seria infinitamente complicada.

Bem, quase. Por muitos anos os historiadores da matemática contaram a história dessa maneira. Mas por volta de 1990, June Barrow-Green descobriu uma cópia impressa do memorial de Poincaré nas profundezas do Instituto Mittag-Leffler em Estocolmo, folheou o exemplar e percebeu que era diferente da versão que podia ser encontrada em numerosas bibliotecas matemáticas ao redor do globo. Era, na verdade, uma impressão oficial do memorial vencedor do prêmio de Poincaré, e nele havia um erro. Quando Poincaré submeteu seu trabalho para o prêmio, ele havia passado por cima das soluções caóticas. Localizou o erro antes que o memorial fosse publicado, elaborou o que deveria ter deduzido – ou seja, caos – e pagou (mais do que o valor do prêmio) para ter a versão original destruída e a versão corrigida impressa. Por algum motivo os arquivos do Instituto Mittag-Leffler mantiveram um exemplar da versão original falha, mas esta ficou esquecida até Barrow-Green desenterrá-la, publicando sua descoberta em 1994.

Poincaré parece ter pensado que essas soluções caóticas eram incompatíveis com expansões de séries, mas isso também se revelou errado. Era uma premissa fácil de se fazer: as séries parecem regulares demais para representar o caos; só a topologia pode fazer isso. O caos é um comportamento complicado causado por regras simples, então a inferência não era inequívoca, mas a estrutura do problema dos três corpos decididamente exclui soluções simples do tipo que Newton deduziu para

dois corpos. O problema dos dois corpos é “integrável”, o que significava que as equações têm grandezas conservadas suficientes, tais como energia, quantidade de movimento e momento angular para determinar as órbitas. “Conservadas” quer dizer que essas grandezas não se alteram à medida que os corpos percorrem suas órbitas. O problema dos três corpos é conhecido por não ser integrável.

Mesmo assim, soluções em termos de séries existem, sim, mas não são válidas universalmente. Elas falham em estados iniciais de momento angular zero – uma medida do giro total – que são infinitamente raros porque zero é apenas um número em meio à infinidade de todos os números reais. Ademais, não há séries no tempo variáveis como tais: são séries em sua raiz cúbica. O matemático finlandês Karl Fritiof Sundman descobriu tudo isso em 1912. Algo similar chega a valer para o problema dos *n* corpos, mais uma vez com raras exceções, resultado obtido em 1991 por Qiudong Wang. Mas para quatro ou mais corpos, não temos classificação alguma das circunstâncias precisas nas quais as séries deixam de convergir. Sabemos que tal classificação deve ser muito complicada, porque existem soluções nas quais todos os corpos escapam para o infinito, ou oscilam com rapidez infinita, após um tempo finito (Capítulo 12). Fisicamente, essas soluções são artefatos da premissa de que os corpos são pontos isolados com massa. Matematicamente, nos dizem onde procurar por comportamento excêntrico.

UM PROGRESSO ESPETACULAR no problema dos *n* corpos foi feito nos casos em que todos os corpos têm a mesma massa. Raramente essa é uma premissa realista em mecânica celeste, mas faz sentido para alguns modelos não quânticos de partículas elementares. O principal interesse é matemático. Em 1993, Christopher Moore encontrou uma solução para o problema dos três corpos na qual todos eles fazem um jogo de “siga-o-líder” ao longo da mesma órbita. Ainda mais surpreendente é o formato da órbita: um número oito, mostrado na Figura 32. Embora a órbita cruze sobre si mesma, os corpos nunca colidem. O cálculo de Moore foi numérico, em um computador. Sua solução foi redescoberta de forma independente em 2001 por Alain Chenciner e Richard Montgomery, que combinaram um antigo princípio válido da mecânica clássica, conhecido como “ação mínima”, com um pouco de topologia distintamente sofisticada para dar uma prova rigorosa de que tal solução existe. As órbitas são periódicas no tempo: após um intervalo de tempo fixo os corpos retornam às suas posições e

velocidades iniciais, e daí em diante repetem indefinidamente os mesmos movimentos. Para uma determinada massa comum, há pelo menos uma solução dessas para qualquer período.

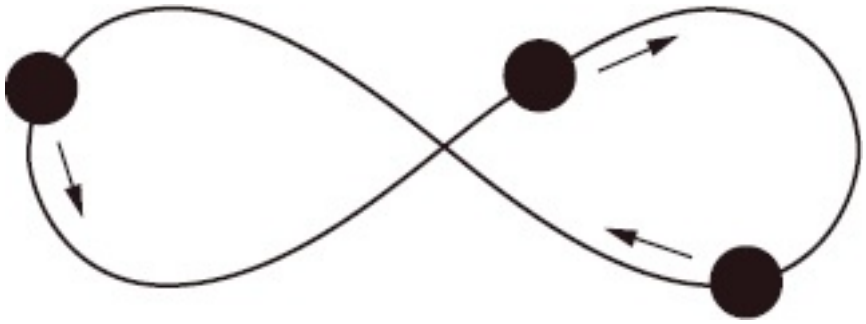


FIGURA 32 A coreografia do número oito.

Em 2000, Carles Simó executou uma análise numérica indicando que o número oito é estável, com exceção talvez de um deslocamento de longo prazo muito lento conhecido como difusão de Arnold, relacionado com a geometria detalhada do mapa de retorno de Poincaré. Para esse tipo de estabilidade, *quase* todas as perturbações levam a uma órbita muito próxima da órbita referida, e à medida que a perturbação torna-se menor, a proporção de tais perturbações aproxima-se de 100%. Para a pequena proporção de perturbações que não se comportam dessa maneira estável, a órbita desloca-se de seu ponto original com extrema lentidão. O resultado de Simó foi uma surpresa, pois órbitas estáveis são raras no problema dos três corpos de mesma massa. Cálculos numéricos mostram que a estabilidade persiste, mesmo quando as três massas são ligeiramente diferentes. Logo, é possível que em algum lugar do universo três estrelas com massas quase idênticas estejam se caçando mutuamente formando o número oito. Em 2000, Douglas Heggie estimou que o número de tais estrelas triplas situa-se em algum ponto entre um por galáxia e um por universo.

O número oito tem uma simetria interessante. Comece com três corpos A, B e C. Siga-os por um terço do período orbital. Você encontrará então os três corpos com as mesmas posições e velocidades que tinham no começo – mas agora os corpos correspondentes são B, C e A. Após dois terços do período, ocorre o mesmo para C, A e B. Um período completo restaura os rótulos originais dos corpos. Esse tipo de solução é conhecido como coreografia: uma dança planetária na qual todo mundo troca de posição de vez em quando. Evidência numérica revela a existência de coreografias

para mais de três corpos: a Figura 33 mostra alguns exemplos. Simó, em particular, descobriu uma quantidade enorme de coreografias.[3]

Mesmo aqui, muitas perguntas permaneceram sem resposta. Carecemos de provas rigorosas a respeito da existência dessas coreografias. Para mais de três corpos todas parecem ser instáveis; muito provavelmente isso está correto, mas precisa ser provado. A órbita em formato de número oito para três corpos de determinada massa e determinado período parece ser especial, porém, mais uma vez, nenhuma prova é conhecida, embora em 2003 Tomasz Kapela e Piotr Zgliczynski tenham fornecido uma prova, com auxílio de computador, de que ela é localmente única – nenhuma órbita próxima funciona. Coreografias podem ser outro grande problema em formação.

E ENTÃO, o sistema solar é estável?

Talvez sim, talvez não.

Dando sequência à grande percepção de Poincaré, a possibilidade do caos, compreendemos agora com muito mais clareza as questões teóricas envolvidas em estabelecer a estabilidade. Elas se revelaram sutis, complexas e – ironicamente – não relacionadas com as soluções de séries de nenhum modo proveitoso. O trabalho de Jürgen Moser e Vladimir Arnold levou a provas de que vários modelos simplificados do sistema solar são estáveis para quase todos os estados iniciais, exceto talvez pelo efeito da difusão de Arnold, que impede tipos mais fortes de estabilidade em quase todos os problemas desse tipo. Em 1961, Arnold provou que, nesse sentido, um modelo idealizado de sistema solar é estável, mas apenas com a premissa de que os planetas tenham massas muito pequenas em comparação com a massa da estrela central, e que as órbitas sejam muito próximas de circulares e muito próximas de um plano comum. No que diz respeito à prova rigorosa, "muito próximo" aqui significa "diferindo por um fator de no máximo 10^{-43}, e, mesmo assim, o enunciado completo é que a probabilidade de ser instável é zero. Nesse tipo de discussão sobre perturbação, os resultados são muitas vezes válidos para discrepâncias maiores do que algo que possa ser provado de maneira rigorosa, portanto a inferência é que sistemas planetários razoavelmente próximos desse ideal são provavelmente estáveis. No entanto, em nosso sistema solar os números relevantes são em torno de 10^{-3} para as massas e 10^{-2} para circularidade e inclinação. Esses valores excedem confortavelmente 10^{-43}.

Assim, a probabilidade de aplicação do resultado de Arnold era discutível. Foi, não obstante, encorajador que *alguma coisa* pudesse ser dita com certeza.

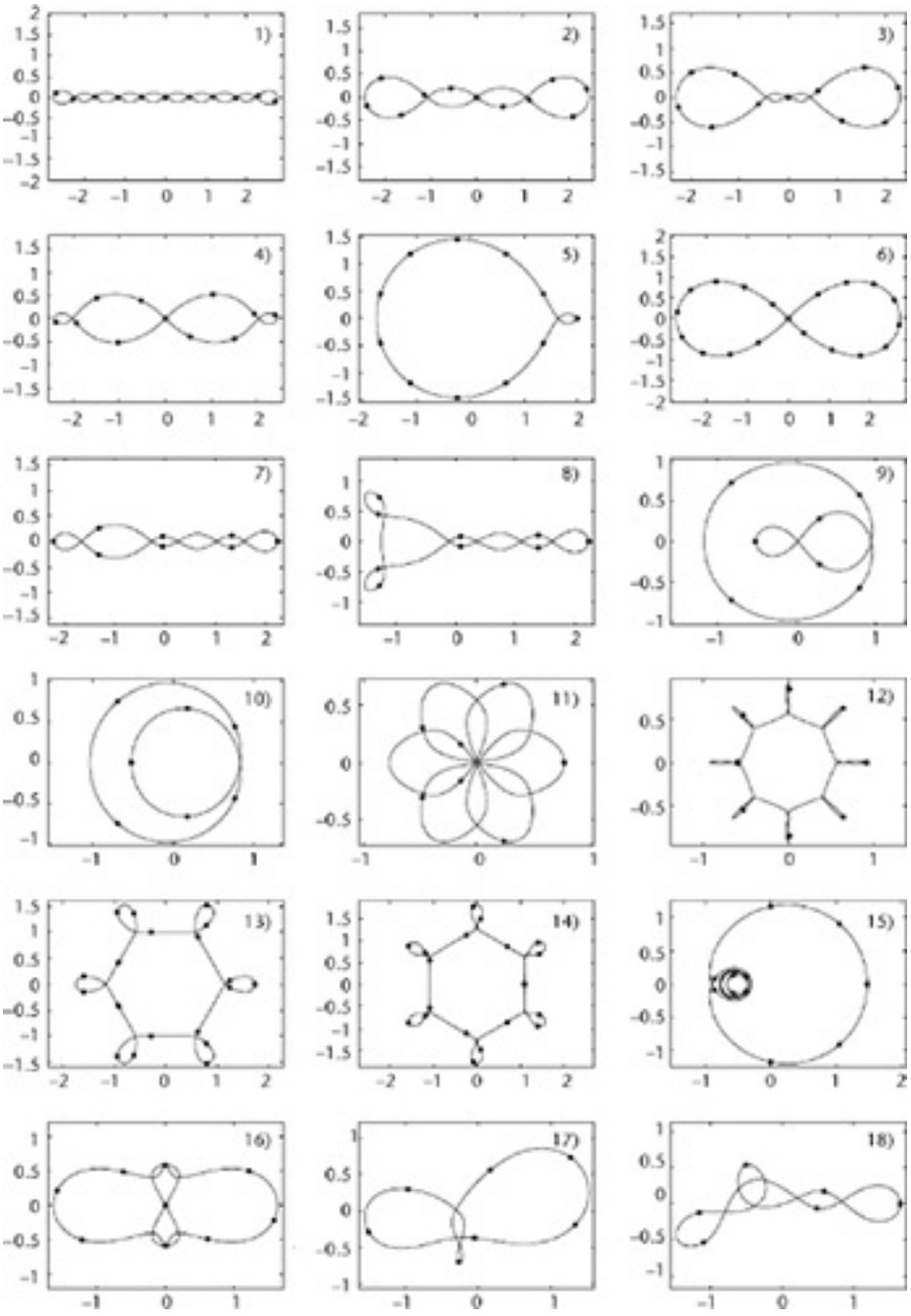


FIGURA 33 Exemplos de coreografias.

As questões práticas em tais problemas também têm ficado mais claras, graças ao desenvolvimento de poderosos métodos numéricos para resolver equações por aproximação pelo computador. Esse é um assunto delicado porque o caos tem uma consequência importante: pequenos erros podem crescer muito rápido e arruinar as respostas. Nossa compreensão teórica

do caos, e de equações como essas para o sistema solar, onde não há atrito, levaram ao desenvolvimento de métodos numéricos que são imunes a muitas das características mais importunas do caos. São chamados de integradores simpléticos. Ao utilizá-los, fica claro que a órbita de Plutão é caótica. No entanto, isso não implica que Plutão corra atropelando o sistema solar, causando tumultos. Significa que em 200 milhões de anos Plutão ainda estará em algum lugar próximo a sua órbita atual, mas não temos a menor pista do paradeiro desse local.

Em 1982, o Projeto Longstop de Archie Roy modelou os planetas externos (Júpiter externo) em um supercomputador, que não encontrou, em larga escala, nenhuma instabilidade, embora alguns dos planetas adquirissem, de maneira estranha, energia à custa de outros. Desde então, sobretudo dois grupos de pesquisa, dirigidos por Jack Wisdom e Jacques Laskar, desenvolveram esses métodos computacionais e os aplicaram a muitos problemas distintos em relação ao nosso sistema solar. Em 1984, o grupo de Wisdom predisse que o satélite de Saturno, Hipérion, em vez de girar com regularidade, deveria sacudir-se caoticamente, e observações subsequentes confirmaram esse fato. Em 1988, em colaboração com Gerry Sussman, o grupo construiu seu próprio computador, talhado para equações de mecânica celeste: o planetário digital. Um planetário é um dispositivo mecânico, com rodas dentadas e engrenagens, capaz de simular o movimento dos planetas, que são pequenas bolas metálicas encaixadas em varetas.[4] O cálculo computacional original acompanhava os próximos 845 milhões de anos do sistema solar, revelando a natureza caótica de Plutão. Com seus sucessores, o grupo de Wisdom tem explorado a dinâmica do sistema solar pelos próximos bilhões de anos.

O grupo de Laskar publicou seus primeiros resultados sobre o comportamento de longo prazo do sistema solar em 1989, usando uma forma que poderia ser considerada uma média das equações e que remonta a Lagrange. Aqui alguns dos detalhes finos são excluídos ou ignorados. Os cálculos do grupo mostraram que a posição da Terra em sua órbita é caótica, bastante parecida com a de Plutão: se medirmos onde a Terra está hoje, e errarmos por quinze metros, então sua posição em órbita daqui a 100 milhões de anos não pode ser prevista com qualquer grau de certeza.

Um meio de mitigar os efeitos do caos é realizar muitas simulações, com dados iniciais ligeiramente diferentes, e obter uma imagem da gama de possíveis futuros e qual a probabilidade de cada um deles. Em 2009, Laskar e Mickaël Gastineau aplicaram essa técnica ao sistema solar,

acompanhando 2.500 cenários diferentes. As diferenças são extraordinariamente pequenas – mover Mercúrio um metro, por exemplo. Em cerca de 1% desses futuros, Mercúrio torna-se instável: colide com Vênus, mergulha no Sol ou é lançado no espaço exterior.

Em 1999, Norman Murray e Matthew Holman investigaram a inconsistência entre resultados como os de Arnold, indicando estabilidade, e as simulações, indicando instabilidade. “Estarão os resultados numéricos incorretos, ou serão os cálculos clássicos simplesmente inaplicáveis?”, eles indagaram. Utilizando métodos analíticos, não numéricos, demonstraram que os cálculos clássicos não se aplicam. As perturbações necessárias para refletir a realidade são grandes demais. A principal fonte de caos no sistema solar é uma quase ressonância entre Júpiter, Saturno e Urano, além de outra menos importante envolvendo Saturno, Urano e Netuno. Também empregaram métodos numéricos para verificar esta argumentação, mostrando que o horizonte de predições – uma medida de tempo que leva pequenos erros a se tornarem grandes o bastante para ter um efeito significativo – é de cerca de 10 milhões de anos.[5] Suas simulações mostram que Urano passa por quase encontros ocasionais com Saturno, à medida que a excentricidade na sua órbita varia de maneira caótica, e há uma chance de que ele eventualmente venha a ser ejetado por completo do sistema solar. No entanto, o tempo provável para que isso ocorra é de cerca de 10^{18} anos. O Sol vai se expandir numa gigante vermelha muito antes disso, de agora a 5 bilhões de anos, o que afetará todos os planetas, no mínimo porque o Sol perderá 30% de sua massa. A Terra se moverá para fora, e poderá escapar por pouco de ser engolfada pelo Sol extremamente expandido. Contudo, hoje acredita-se que as interações das marés acabem por puxar a Terra para o Sol. Os oceanos terrestres já terão fervido e evaporado antes disso. Mas, considerando que o tempo de vida típico de uma espécie, em termos evolucionários, não passa de 5 milhões de anos, nós realmente não precisamos nos preocupar com qualquer uma dessas catástrofes potenciais. Alguma outra coisa vai nos levar primeiro.

Os mesmos métodos podem ser utilizados para investigar o passado do sistema solar: usam-se as mesmas equações e simplesmente faz-se o tempo correr ao contrário, um truque matemático simples. Até pouco tempo, os astrônomos tinham a tendência de assumir que os planetas sempre estiveram perto de suas órbitas atuais, desde que se condensaram a partir de uma nuvem de gás e poeira em volta do nascente Sol. Na verdade, suas órbitas e composições têm sido usadas para inferir o tamanho e a

composição dessa nuvem de poeira primeva. Agora parece que os planetas não começaram em suas órbitas presentes. À medida que a nuvem de poeira coalesceu, sob suas próprias forças gravitacionais, Júpiter – o planeta de maior massa – começou a organizar as posições dos outros corpos, e estes por sua vez influenciaram-se mutuamente. Essa possibilidade foi proposta em 1984 por Julio Fernandez e Wing-Huen Ip, mas por algum tempo seu trabalho foi encarado como uma curiosidade menor. Em 1993, Renu Malhotra começou a pensar seriamente na maneira como as variações na órbita de Netuno poderiam influenciar os outros planetas gigantes; outros entraram no jogo, e começou a surgir a imagem de um sistema solar inicial bastante dinâmico.

À medida que os planetas continuaram a se agregar, chegou uma hora em que Júpiter, Saturno, Urano e Netuno estavam quase completos, mas entre eles circulavam enormes quantidades de planetesimais rochosos e gelados, pequenos corpos com cerca de dez quilômetros de diâmetro. Desse ponto em diante, o sistema solar evoluiu por meio da migração e colisão dos planetesimais. Muitos foram ejetados, o que reduziu a energia e o momento angular dos quatro planetas gigantes. Uma vez que esses mundos tinham diferentes massas e estavam a diferentes distâncias do Sol, reagiram de formas distintas. Netuno foi um dos vencedores nas apostas de energia orbital, e migrou para fora. O mesmo aconteceu com Urano e Saturno, em menor medida. Júpiter foi o grande perdedor em termos de energia, e moveu-se para dentro. Mas era tão pesado que não foi muito longe.

Os outros corpos menores do sistema solar também foram afetados por essas mudanças. O plano do nosso sistema solar atual, aparentemente estável, surgiu mediante uma intrincada dança dos gigantes, na qual eles jogaram os corpos menores uns contra os outros num tumultuado caos. E então, o sistema solar é estável? Provavelmente não, mas não estaremos mais aqui para descobrir.

9. Padrões em primos

A hipótese de Riemann

NO CAPÍTULO 2 examinamos as propriedades dos números primos como indivíduos, e eu as comparei ao comportamento muitas vezes errático e imprevisível dos seres humanos. Homens têm livre-arbítrio; podem fazer suas próprias escolhas por suas próprias razões. Os primos têm a ver com o que quer que a lógica da aritmética lhes imponha, mas com frequência também parecem ter seu próprio arbítrio. Seu comportamento é governado por estranhas coincidências e muitas vezes carece de qualquer estrutura coerente.

Entretanto, o mundo dos números primos não é regido por anarquia. Em 1835, Adolphe Quetelet estarreceu seus contemporâneos encontrando regularidades matemáticas genuínas em eventos sociais que dependiam de escolhas humanas conscientes ou da intervenção do destino: nascimentos, casamentos, mortes, suicídios. Os padrões eram estatísticos: referiam-se não a indivíduos, mas a um comportamento médio de grande número de pessoas. É assim que os estatísticos extraem ordem do livre-arbítrio individual. Praticamente na mesma época, os matemáticos começaram a perceber que o mesmo truque funciona com os primos. Embora cada primo seja um individualista inveterado, coletivamente eles se conformam à regência da lei. Existem padrões ocultos.

Padrões estatísticos surgem quando pensamos em intervalos inteiros de números primos. Por exemplo: quantos primos existem até um limite específico? É uma pergunta muito difícil de ser respondida com exatidão, mas há aproximações excelentes, e quanto maior o limite, melhores se tornam as aproximações. Às vezes a diferença entre a aproximação e a resposta exata pode se tornar muito pequena, mas isso já é pedir demais. A maioria das aproximações nessa área é assintótica, o que significa que a razão entre a aproximação e a resposta exata pode se tornar muito próxima de 1. O erro absoluto na aproximação pode crescer até atingir qualquer tamanho, mesmo que o erro percentual diminua tendendo a zero.

Se você está se perguntando como isso é possível, suponha que a sequência aproximada de números para alguma propriedade obtusa de

primos sejam as potências de 100:

100 10.000 1.000.000 100.000.000

mas os números exatos sejam

101 10.010 1.000.100 100.001.000

onde o 1 extra move-se uma casa para a esquerda em cada estágio. Então as razões dos números correspondentes aproximam-se mais e mais de 1, mas as diferenças são

1 10 100 1.000

tornando-se tão grandes quanto desejarmos. Esse tipo de comportamento ocorre se os erros – as diferenças entre a aproximação e a resposta exata – crescerem sem limite, mas aumentarem mais lentamente que os números em si.

A busca por fórmulas assintóticas relacionadas com os primos inspirou novos métodos na teoria dos números, baseados não em números inteiros, mas em análise complexa. Análise é uma formulação rigorosa do cálculo, que tem dois aspectos fundamentais. Um, o cálculo diferencial, trata da taxa em que certa grandeza, chamada função, varia em relação a outra grandeza. Por exemplo, a posição de um corpo depende – é função – do tempo, e a taxa de variação dessa posição à medida que o tempo passa é a velocidade instantânea do corpo. O outro aspecto, o cálculo integral, trata do cálculo de áreas, volumes e similares somando-se grandes quantidades de pedaços muito pequenos, processo este chamado integração. Notavelmente, a integração acaba por ser o inverso da diferenciação. A formulação original do cálculo feita por Newton e Gottfried Leibniz exigia algumas manobras com grandezas infinitamente pequenas, levantando questões a respeito da validade lógica da teoria. Essas questões conceituais foram resolvidas definindo-se a noção de limite, um valor do qual é possível aproximar-se quanto se queira, mas que não precisa necessariamente ser atingido. Quando apresentada em sua forma mais rigorosa, o tópico é chamado análise.

No tempo de Newton e Leibniz, as grandezas envolvidas eram números reais, e a matéria que emergiu foi a análise real. À medida que os números complexos foram se tornando amplamente aceitos entre os matemáticos, foi natural estender a análise a quantidades complexas. Essa é a matéria da

análise complexa, que acabou revelando-se extraordinariamente bela e poderosa. Quando se trata de análise, as funções complexas têm um comportamento muito melhor que as funções reais. Elas possuem suas peculiaridades, é claro, mas as vantagens de trabalhar com funções complexas supera bastante as desvantagens.

Foi uma grande surpresa quando os matemáticos descobriram que as características aritméticas dos números inteiros podem ser vantajosamente reformuladas em termos de funções complexas. Antes, esses dois sistemas de números formulavam perguntas muito diferentes e empregavam métodos muito distintos. Mas agora a análise complexa, um corpo de técnicas extraordinariamente poderosas, podia ser usada para descobrir características especiais de funções na teoria dos números; e a partir delas, podiam-se extrair fórmulas assintóticas e muitas coisas mais.

Em 1859, um matemático alemão, Bernhard Riemann, lançou mão de uma antiga ideia de Euler e desenvolveu-a de uma maneira drasticamente nova, definindo a chamada função zeta. Uma das consequências foi uma fórmula *exata* para o número de primos até certo limite. Era uma soma infinita, mas os analistas estavam acostumados com elas. Não se tratava apenas de um artifício esperto mas inútil; fornecia percepções genuinamente novas dos primos. Havia apenas um pequeno obstáculo. Embora Riemann pudesse provar que sua fórmula era exata, suas mais importantes consequências potenciais dependiam de um enunciado simples sobre a função zeta, e esse enunciado Riemann não conseguiu provar. Um século e meio depois, ainda não conseguimos. Ela é chamada hipótese de Riemann, e é o Santo Graal da matemática pura.

NO CAPÍTULO 2 vimos que os primos tendem a rarear à medida que vão ficando maiores. Considerando que resultados exatos sobre sua distribuição pareciam fora de cogitação, por que não procurar em vez disso padrões estatísticos? Em 1797-8, Legendre contou quantos primos ocorrem até vários limites, usando tabelas de primos que haviam sido recentemente fornecidas por Jurij Vega e Anton Felkel. Vega devia gostar de cálculos longos: construiu tabelas de logaritmos e em 1789 foi detentor do recorde mundial para calcular π, até 140 casas decimais (126 corretas). Felkel simplesmente gostava de calcular números primos. Sua principal obra, de 1776, é *Tafel aller Einfachen Factoren der durch 2, 3, 5 nicht theilbaren Zahlen von 1 bis 10 000 000* (Tabela de todos os fatores comuns

indivisíveis por 2, 3, 5 dos números de 1 a 10.000.000). Há maneiras fáceis de testar os fatores 2, 3 e 5, mencionados no Capítulo 2, de modo que ele economizou muito espaço omitindo esses números. Legendre descobriu uma aproximação empírica para a quantidade de primos menores que um número dado x, que é representada por $\pi(x)$. Se você conhece π apenas como um símbolo para o número 3,14159, demora um pouco para se acostumar, mas não é difícil perceber o que se pretende, mesmo sem notar que as fontes são diferentes. O texto de Legendre de 1808 sobre a teoria dos números afirma que $\pi(x)$ parece estar bastante próximo de $x/(\ln x - 1{,}08366)$.

Em uma carta de 1849 para o astrônomo Johann Encke, Gauss afirmou que quando tinha cerca de quinze anos escreveu uma anotação em suas tábuas de logaritmos, dizendo que o número de primos menores ou iguais a x é $x/\ln x$ para um x grande. Como em muitas outras de suas descobertas, Gauss não publicou essa aproximação, talvez porque não tivesse prova. Em 1838, Dirichlet informou a Gauss uma aproximação similar que havia descoberto, que se resume na função integral logarítmica[1]

$$\mathrm{Li}(x) = \int_0^x \frac{dt}{\ln t}$$

A razão de $\mathrm{Li}(x)$ para $x/\ln x$ tende a 1 à medida que x torna-se grande, implicando que se um é assintótico a $\pi(x)$ o outro também é, mas a Figura 34 sugere (corretamente) que $\mathrm{Li}(x)$ é uma aproximação melhor do que $x/\ln x$. A precisão de $\mathrm{Li}(x)$ é bastante impressionante; por exemplo:

$\pi(1.000.000.000) = 50.847.534$
$\mathrm{Li}(1.000.000.000) = 50.849.234{,}9$

A de $x/\ln x$ é pior: aqui o resultado é 48.254.942,4.

A fórmula de aproximação – seja usando $\mathrm{Li}(x)$ ou $x/\ln x$ – tornou-se conhecida como teorema dos números primos, onde a palavra "teorema" foi usada no sentido de "conjectura". A busca por uma prova de que essas fórmulas são assintóticas a $\pi(x)$ tornou-se um dos principais problemas em aberto na teoria dos números. Muitos matemáticos o atacaram usando métodos tradicionais nessa área, e alguns poucos chegaram perto; todavia, sempre parecia haver alguma premissa traiçoeira que impedia a prova. Eram necessários métodos novos. E eles vieram de uma curiosa reformulação de dois dos antigos teoremas de Euclides sobre os números

primos.

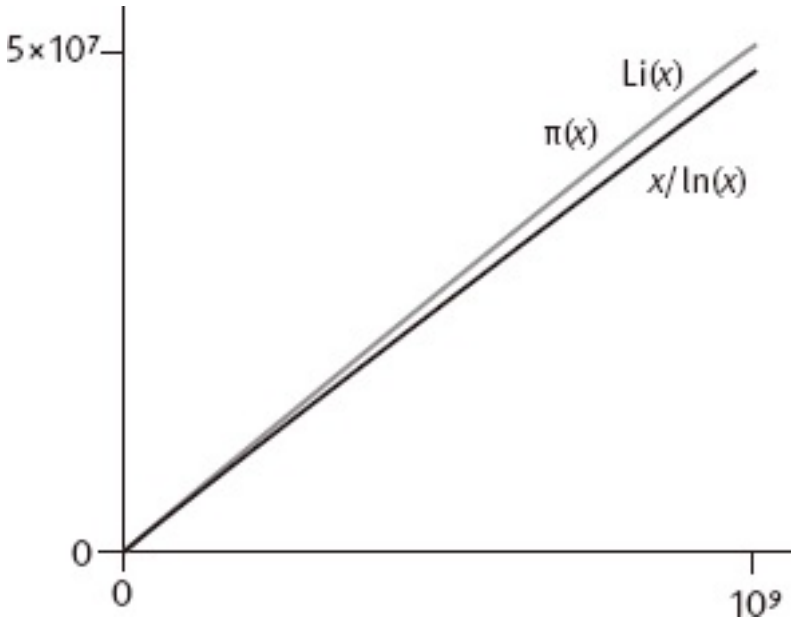


FIGURA 34 Nesta escala, $\pi(x)$ e $\mathrm{Li}(x)$ (cinza) são indistinguíveis. No entanto, $x/\ln x$ (preto) é visivelmente menor. Aqui x corre na horizontal e o valor da função é marcado no eixo vertical.

O TEOREMA DOS NÚMEROS PRIMOS era uma resposta ao teorema de Euclides de que os primos continuam para sempre. Outro teorema básico de Euclides é a exclusividade da decomposição em fatores primos: cada inteiro positivo é produto de primos *exatamente de uma única maneira.* Em 1737, Euler percebeu que o primeiro teorema pode ser apresentado de outro modo, como uma fórmula bastante surpreendente em análise real, e a segunda afirmação torna-se uma simples consequência dessa fórmula. Começarei apresentando a fórmula, depois tentarei dar sentido a ela. Ei-la:

$$\frac{1}{1-2^{-s}} \times \frac{1}{1-3^{-s}} \times \ldots \times \frac{1}{1-p^{-s}} \times \ldots = \frac{1}{1^s} + \frac{1}{2^s} + \frac{1}{3^s} + \frac{1}{4^s} + \frac{1}{5^s} + \frac{1}{6^s} + \frac{1}{7^s} + \ldots$$

Aqui p passa por todos os primos e s é constante. Euler estava particularmente interessado no caso em que s é um número inteiro, mas sua fórmula funciona também para números reais, contanto que s seja maior que 1. Essa condição é necessária para fazer a série do lado direito convergir: ter um valor significativo quando continuada indefinidamente.

Essa é uma fórmula extraordinária. Do lado esquerdo nós multiplicamos entre si infinitas expressões que dependem apenas dos

primos. Do lado direito somamos entre si infinitas expressões que dependem de todos os números inteiros positivos. A fórmula expressa, em linguagem analítica, algumas relações entre números inteiros e primos. A principal relação desse tipo é a exclusividade da decomposição em fatores primos, e é isso que justifica a fórmula.

Esboçarei o passo principal para mostrar que existe uma ideia sensata por trás de tudo isso. Usando a álgebra escolar podemos expandir a expressão em *p* numa série, parecida com a expressão do lado direito da fórmula, mas envolvendo apenas potências de *p*. Especificamente,

$$\frac{1}{1-p^{-s}} = \frac{1}{1^s} + \frac{1}{p^s} + \frac{1}{p^{2s}} + \frac{1}{p^{3s}} + \dots$$

Quando multiplicamos todas essas séries entre si, para todos os primos *p*, e "expandimos" para obter uma soma de termos simples, obtemos toda combinação de potências primas – ou seja, todo número inteiro. Cada um ocorre como o inverso (1 dividido por) de sua potência de ordem *s*, e cada um ocorre exatamente uma vez devido à exclusividade da fatoração em primos. Assim, obtemos a série à direita.

Ninguém jamais achou uma fórmula algébrica simples para essa série, embora haja muitas usando integrais. Assim, nós lhe damos um símbolo especial, a letra grega zeta (ζ), e definimos uma nova função:

$$\zeta(s) = \frac{1}{1^s} + \frac{1}{2^s} + \frac{1}{3^s} + \frac{1}{4^s} + \frac{1}{5^s} + \frac{1}{6^s} + \frac{1}{7^s} + \dots$$

Euler não usou efetivamente o símbolo ζ, e considerou apenas valores inteiros positivos de *s*, mas chamarei a série acima de função zeta de Euler. Usando sua fórmula, Euler deduziu que existe uma infinita quantidade de números primos fazendo *s* chegar muito perto de 1. Se houver uma quantidade finita de primos, o lado esquerdo da fórmula tem um valor finito, mas o lado direito torna-se infinito. Isso é uma contradição, logo deve haver infinitos primos. O principal objetivo de Euler era obter fórmulas como $\zeta(2) = \pi^2/_6$, dando a soma da série para inteiros pares *s*. Ele não levou sua revolucionária ideia muito adiante.

OUTROS MATEMÄTICOS IDENTIFICARAM o que Euler tinha deixado passar, e consideraram os valores de *s* não inteiros. Em dois artigos de 1848 e 1850 o matemático russo Pafnuty Chebyshev teve uma ideia brilhante: tentar

provar o teorema dos números primos usando análise.[2] Ele começou com o elo entre números primos e análise fornecido pela função zeta de Euler. Não foi muito bem-sucedido, porque pressupôs *s* real, e as técnicas analíticas disponíveis na análise real eram limitadas demais. Mas conseguiu provar que quando *x* é grande, a razão $\pi(x)$ para $x/\ln x$ situa-se entre duas constantes: uma ligeiramente maior que 1 e outra ligeiramente menor. Foi uma recompensa genuína, mesmo com esse resultado mais fraco, porque permitiu-lhe provar o postulado de Bertrand, conjecturado em 1845: se você pegar um inteiro qualquer e duplicá-lo, existirá um primo entre ambos.

Agora o palco estava armado para Riemann. Ele também reconheceu que a função zeta contém a chave para o mistério do teorema dos números primos, mas para fazer sua abordagem funcionar teve de propor uma extensão ambiciosa: definir a função zeta não só para uma variável real, mas para uma complexa. A série de Euler é um bom lugar para começar. Ela converge para todo *s* real maior que 1, e se exatamente a mesma fórmula for usada para um *s* complexo, então a série converge sempre que a parte real de *s* for maior que 1. No entanto, Riemann descobriu que podia fazer algo muito melhor. Usando um procedimento chamado continuação analítica, estendeu a definição de $\zeta(s)$ para *todos* os números complexos diferentes de 1. Esse valor de *s* fica excluído porque a função zeta torna-se infinita quando $s = 1$.[3]

Em 1859, Riemann reuniu suas ideias acerca da função zeta em um artigo, cujo título traduzido é “Sobre a quantidade de primos menores que uma magnitude dada”.[4] Nesse artigo, Riemann dava uma fórmula explícita, exata para $\pi(x)$.[5] Vou descrever uma fórmula mais simples, equivalente à de Riemann, para mostrar como os zeros da função zeta aparecem. A ideia é contar quantos primos ou potências de primos existem até um limite escolhido. No entanto, em vez de contar cada um deles uma vez, que é o que $\pi(x)$ faz com os primos, os primos maiores recebem um peso extra. Na verdade, qualquer potência de um número primo é recontada de acordo com o logaritmo desse primo. Por exemplo, até o limite de 12 as potências de primos são

$2, 3, 4 (= 2^2), 5, 7, 8 (= 2^3), 9 (= 3^2), 11$

então, a contagem ponderada fica:

$\ln 2 + \ln 3 + \ln 2 + \ln 5 + \ln 7 + \ln 2 + \ln 3 + \ln 11$

que é aproximadamente 10,23.

Utilizando análise, informação sobre essa maneira mais sofisticada de contar números primos pode ser transformada em informação sobre o modo habitual. Todavia, esse modo leva a fórmulas mais simples, um pequeno preço a pagar pelo uso do logaritmo. Nesses termos, a fórmula exata de Riemann afirma que a contagem ponderada até um limite *x* é igual a onde Σ indica a somatória de todos os números *p* para os quais ζ(*p*) é zero, excluindo inteiros pares negativos. Esses números são chamados de zeros não triviais da função zeta. Os zeros triviais são os inteiros pares negativos −2, −4, −6, … A função zeta é zero nesses valores devido à fórmula usada na definição da continuação analítica, mas esses zeros acabam se revelando não importantes para a fórmula de Riemann, ou para muita outra coisa.

$$-\sum_{\rho} \frac{x^{\rho}}{\rho} + x - \frac{1}{2}\ln(1-x^{-2}) - \ln 2\pi$$

Se por acaso a fórmula parece um pouco intimidadora, deixe-me pinçar o ponto principal: um modo extravagante de contar números primos até um limite *x*, que pode ser transformado no jeito habitual com um pouquinho de artifícios analíticos, é *exatamente* igual a somar todos os zeros não triviais da função zeta da simples expressão $x^{\rho}/_{\rho}$, *mais uma função direta de x*. Se você for um analista complexo, verá imediatamente que o teorema dos números primos é equivalente a provar que a contagem ponderada até o limite *x* é assintótica a *x*. Usando análise complexa, isso será verdade se todos os zeros não triviais da função zeta tiverem partes reais entre 0 e 1. Chebyshev não conseguiu provar isso, mas chegou perto o bastante para obter informação útil.

Por que os zeros da função zeta são tão importantes? Um teorema básico em análise complexa afirma que, sujeita a algumas condições técnicas, uma função de variável complexa é completamente determinada pelos valores da variável para os quais o valor da função é zero ou infinito, junto com alguma informação adicional sobre seu comportamento nesses pontos. Esses lugares especiais são conhecidos como os zeros e polos da função. Esse teorema não funciona em análise real – uma das muitas razões por que a análise complexa tornou-se o cenário preferido, apesar de exigir a raiz quadrada de menos um. A função zeta tem um polo, em *s* = 1, então tudo em relação a ela é determinado pelos seus zeros, com a condição de mantermos em mente esse polo único.

Por conveniência, Riemann trabalhou basicamente com uma função correlata, a função xi $\xi(x)$, que está intimamente relacionada com a função zeta, e emerge a partir do método da continuação analítica. Riemann comentou que:

É muito provável que todos [zeros da função xi] sejam reais. Deve-se, porém, desejar uma prova estrita disso; todavia eu, após algumas efêmeras tentativas fúteis, deixei provisoriamente de lado a busca por essa prova, pois parece desnecessária para o objetivo seguinte da minha investigação.

A afirmação sobre a função xi é equivalente a uma afirmação sobre a correlata função zeta. Isto é, todos os zeros não triviais da função zeta são números complexos da forma $\frac{1}{2} + it$ e estão sobre a *linha crítica* "parte real igual a ½" (Figura 35). Essa versão de seu comentário é a famosa hipótese de Riemann.

O comentário de Riemann é bastante casual, como se sua hipótese não fosse de grande importância. Em relação ao seu programa para provar o teorema dos números primos de fato não era. Mas para muitas outras questões, o inverso é verdadeiro. Na verdade, a hipótese de Riemann é amplamente considerada a mais importante questão não respondida na matemática.

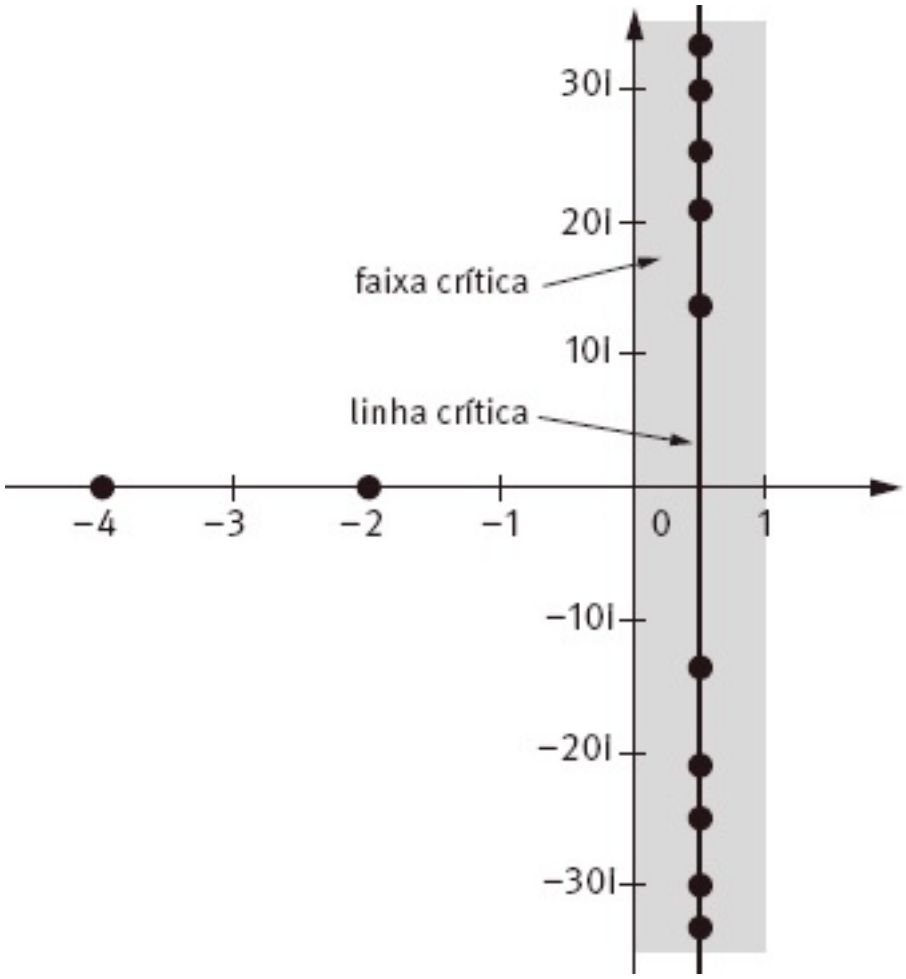


FIGURA 35 Zeros da função zeta, a linha crítica e a faixa crítica.

Para entender por quê, precisamos seguir um pouco mais o raciocínio de Riemann. Ele tinha em vista o teorema dos números primos. Sua fórmula exata sugeria um meio de chegar a isso: entender os zeros da função zeta, ou, de maneira equivalente, da função xi. A hipótese completa de Riemann não é necessária; basta provar que todos os zeros não triviais da função zeta têm partes reais entre 0 e 1. Ou seja, situam-se dentro da distância de ½ da linha crítica de Riemann, na assim chamada faixa crítica. Essa propriedade dos zeros implica que a somatória dos zeros da função zeta, na fórmula exata acima, é uma constante finita. Assintoticamente, para um *x* grande, ela pode muito bem não estar ali. Entre os termos da fórmula, o único que permanece importante à medida que *x* se torna muito grande é o próprio *x*. Todo obstáculo desaparece assintoticamente em comparação a *x*. Portanto, a contagem ponderada é assintótica a *x*, e isso prova o teorema dos números primos. Logo, ironicamente, o papel dos zeros na função zeta é provar que os zeros da função zeta não têm contribuição significativa para a fórmula exata.

Riemann nunca conduziu esse programa até uma conclusão. Na

verdade, nunca mais escreveu sobre o tópico. Mas dois outros matemáticos assumiram o desafio, e mostraram que o palpite de Riemann estava correto. Em 1896, Jacques Hadamard e Charles Jean de la Vallée Poussin deduziram de maneira independente o teorema dos números primos provando que todos os zeros não triviais da função zeta estão na faixa crítica. Suas provas eram muito complicadas e técnicas; mesmo assim, funcionavam. Uma nova e poderosa área da matemática passou a existir: a teoria analítica dos números, que tinha aplicações por toda a teoria dos números, solucionando velhos problemas e revelando novos padrões. Outros matemáticos descobriram provas analíticas mais simples do teorema dos números primos, e Atle Selberg e Paul Erdös descobriram uma prova muito complicada que não requeria análise complexa alguma. Mas a essa altura a ideia de Riemann já fora usada para provar inúmeros teoremas importantes, inclusive aproximações a muitas funções da teoria dos números. Assim, essa nova prova acrescentou uma nota de rodapé irônica, mas teve pouco efeito. Em 1980, Donald Newman descobriu uma prova muito mais simples, usando apenas um dos resultados mais básicos da análise complexa, conhecida como teorema de Cauchy.

EMBORA RIEMANN declarasse sua hipótese como desnecessária para seus objetivos imediatos, ela acabou revelando-se vital para muitas outras questões em teoria dos números. Antes de discutir a hipótese de Riemann, vale a pena dar uma olhada em alguns dos teoremas que se seguiriam caso a hipótese pudesse ser provada verdadeira.

Uma das implicações mais importantes é o tamanho do erro no teorema dos números primos. O teorema afirma que para um x grande a razão entre $\pi(x)$ e $\mathrm{Li}(x)$ vai se aproximando cada vez mais de 1. Ou seja, o tamanho da diferença entre essas duas funções tende a zero, *em relação ao tamanho de x.* No entanto, a diferença real pode ficar cada vez maior (e fica). Ela simplesmente aumenta numa taxa de variação menor do que o aumento do próprio x.[6] Experimentos de computador sugerem que o tamanho do erro é aproximadamente proporcional a $\sqrt{x}\ln x$. Se a hipótese de Riemann for verdadeira, essa afirmação pode ser provada. Em 1901, Helge von Koch provou que a hipótese de Riemann é logicamente equivalente à estimativa

$$\left|\pi(x) - \mathrm{Li}(x)\right| \leq \frac{1}{8\pi}\sqrt{x}\ln x$$

para todo $x \geq 2.657$. As barras verticais | | indicam o valor absoluto: a diferença multiplicada por ± 1 para torná-la positiva. Essa fórmula fornece o melhor limite possível para a diferença entre $\pi(x)$ e $\mathrm{Li}(x)$.

A hipótese de Riemann implica muitas estimativas para outras funções na teoria dos números. Por exemplo, ela é equivalente à soma dos divisores de *n* ser inferior a

$$e^{\gamma} n \ln \ln n$$

para todo $n \geq 5.040$, onde $\gamma = 0{,}57721\ldots$ é a constante de Euler.[7] Esses fatos parecem esquisitices, mas boas estimativas para funções importantes são essenciais para muitas aplicações, e a maioria dos teóricos de números daria o braço direito para provar qualquer uma delas.

A hipótese de Riemann também nos diz de que tamanho pode ser o espaçamento entre dois primos consecutivos. Podemos deduzir o tamanho característico desse espaço a partir do teorema dos números primos: em média, o espaço entre um primo *p* e o seguinte é comparável a $\ln p$. Alguns intervalos são menores, outros maiores, e a vida dos matemáticos ficaria bem mais fácil caso soubessem de que tamanho o espaçamento pode ficar. Harald Cramér provou em 1936 que se a hipótese de Riemann for verdadeira, o espaço num primo *p* não é maior do que uma constante vezes $\sqrt{p} \ln p$.

O VERDADEIRO SIGNIFICADO da hipótese de Riemann reside muito mais fundo. Há inúmeras generalizações, além de um forte palpite de que aquele que conseguir provar a hipótese de Riemann poderá provavelmente provar a hipótese correspondente generalizada de Riemann. Que por sua vez daria aos matemáticos um bocado de controle sobre amplas áreas da teoria dos números.

A hipótese de Riemann generalizada surge a partir de uma descrição mais refinada dos números primos. Todos os primos exceto 2 são ímpares, e vimos no Capítulo 2 que os ímpares podem ser classificados em dois tipos: os que têm uma unidade a mais que um múltiplo de 4 e os que têm 3 unidades a mais que um múltiplo de 4. Diz-se que são da forma $4k + 1$ e $4k + 3$, onde *k* é o número pelo qual se multiplica 4 para obtê-los. Eis uma pequena lista dos primeiros primos de cada tipo, junto com os correspondentes múltiplos de 4.

múltiplo de 4	0	4	8	12	16	20	24	28	32	36
mais 1	·	5	·	13	17	·	·	29	·	37
mais 3	3	7	11	·	19	23	·	·	·	·

O ponto indica que o referido número não é primo.

Quantos números primos há de cada tipo? Como estão distribuídos entre os primos, ou entre todos os inteiros? A prova de Euclides de que existem infinitos números primos pode ser modificada, sem muito esforço, para provar que existem infinitos primos da forma $4k + 3$. É muito mais difícil provar que existem infinitos primos da forma $4k + 1$; pode ser feito, mas só usando alguns teoremas bastante difíceis. A diferença surge porque qualquer número da forma $4k + 3$ tem algum fator dessa forma; o mesmo nem sempre é verdadeiro para números do tipo $4k + 1$.

Aqui não há nada de sagrado em relação aos números. Exceto 2 e 3, todos os primos são ou da forma $6k + 1$ ou $6k + 5$, e podemos fazer perguntas semelhantes. Quanto a isso, todos os primos, exceto 5, assumem uma das formas $5k + 1$, $5k + 2$, $5k + 3$, $5k + 4$. Deixamos $5k$ de fora porque são múltiplos de 5, então todos eles, com exceção de 5, não são primos.

Não é difícil surgir com uma adivinhação sensata para todas as perguntas desse tipo – primos em sequência aritmética. O caso $5k$ é típico. O experimento rapidamente sugere que números dos quatro tipos listados acima têm bastante chance de serem primos. Eis uma tabela semelhante:

múltiplo de 5	5	10	15	20	25	30	35	40
mais 1	·	11	·	·	·	31	·	41
mais 2	7	·	17	·	·	·	37	·
mais 3	·	13	·	23	·	·	·	43
mais 4	·	·	19	·	29	·	·	·

Logo, deve haver infinitos de cada tipo específico, e em média cerca de um quarto dos primos, até certo limite dado, devem ser de uma forma específica.

Provas simples mostram que algumas formas levam a infinitos números primos, provas mais sofisticadas funcionam para outras formas, mas até meados do século XIX ninguém conseguiu provar que existem infinitos primos de cada forma possível, muito menos que as proporções são aproximadamente iguais. Lagrange assumiu isso sem prova alguma em seu

trabalho a respeito da lei da reciprocidade quadrática – uma propriedade profunda dos quadrados para um módulo primo – em 1785. Os resultados tiveram claramente consequências úteis, e estava mais do que na hora de alguém prová-los. Em 1837, Dirichlet descobriu como adaptar as ideias de Riemann sobre o teorema dos números primos para provar ambas essas afirmações. O primeiro passo foi definir análogas à função zeta para esses tipos de primos. As funções resultantes são chamadas funções-*L* de Dirichlet. Um exemplo, proveniente do caso $^{4k+1}/_{4k+3}$, é

$$L(s, \chi) = 1 - 3^{-s} + 5^{-s} - 7^{-s} + 9^{-s} - \ldots$$

onde os coeficientes são +1 para números da forma $4k + 1$, −1 para números $4k + 3$ e 0 para o restante. A letra grega χ é chamada caractere de Dirichlet, e nos lembra de usar esses signos.

Para a função zeta de Riemann o que importa não é somente a série, mas sua continuação analítica, que dá à função um significado para todos os números complexos. O mesmo vale para a função-*L*, e Dirichlet definiu uma continuação analítica conveniente. Adaptando a ideia usada para provar o teorema dos números primos, ele foi então capaz de provar um teorema análogo para primos de formas específicas. Por exemplo, a quantidade de primos da forma $5k + 1$ menores ou iguais a x é assintótica a $^{\mathrm{Li}(x)}/_{4}$, e o mesmo vale para os outros três casos $5k + 2$, $5k + 3$ e $5k + 4$. Em particular, há infinitos primos de cada forma.

A função zeta de Riemann é um caso especial de uma função-*L* de Dirichlet para primos da forma $1k + 0$, ou seja, todos os primos. A hipótese generalizada de Riemann é a generalização óbvia da hipótese de Riemann original: os zeros de qualquer função-*L* de Dirichlet ou têm parte real ½, ou são "zeros triviais", com parte real ou negativa ou maior que 1.

Se a hipótese generalizada de Riemann for verdadeira, então a hipótese de Riemann também será. Muitas das consequências da hipótese generalizada de Riemann são análogas às da hipótese original. Por exemplo, limites de erros similares podem ser provados para versões análogas do teorema dos números primos, aplicadas a primos de qualquer forma específica. Contudo, a hipótese generalizada de Riemann implica muitos aspectos que são bem diferentes de qualquer coisa que possamos deduzir usando a hipótese de Riemann comum. Assim, em 1917, Godfrey Harold Hardy e John Edensor Littlewood provaram que a hipótese de Riemann generalizada implica uma conjectura de Chebyshev, para efeito de

que (num sentido preciso) números primos da forma $4k + 3$ são mais comuns do que os da forma $4k + 1$. Ambos os tipos são igualmente prováveis, em longo prazo, pelo teorema de Dirichlet, mas isso não impede os primos $4k + 3$ de superarem em competição os primos $4k + 1$ caso se estabeleça o jogo correto.

A hipótese generalizada de Riemann tem também importantes implicações para testes de primalidade, tais como o teste de Miller de 1976 mencionado no Capítulo 2. Se a hipótese generalizada de Riemann for verdadeira, então o teste de Miller fornece um algoritmo eficiente. Estimativas da eficiência de testes mais recentes dependem, do mesmo modo, da hipótese generalizada de Riemann. Existem aplicações significativas também para a teoria algébrica dos números. Lembremo-nos do Capítulo 7, de que a reformulação feita por Dedekind dos números ideais de Kummer levou a um conceito novo e fundamental, os ideais. A fatoração em primos nos anéis de inteiros algébricos existe, mas pode não ser exclusiva. A fatoração em números primos de ideais é muito mais precisa: são válidas tanto a existência quanto a exclusividade. Então, faz sentido reinterpretar todas as questões sobre fatores em termos de ideais. Em especial, há uma noção de "ideal primo", uma analogia razoável e tratável de um número primo.

Sabendo disso, é natural perguntar se o elo estabelecido por Euler entre números primos comuns e a função zeta têm alguma analogia para ideais primos. Se tiver, todo o poderoso maquinário da teoria analítica dos números torna-se disponível para números algébricos. O resultado é a função zeta de Dedekind: uma função dessas para cada sistema de números algébricos. Há uma profunda ligação entre as propriedades analíticas complexas da função zeta de Dedekind e a aritmética dos números primos para os correspondentes inteiros algébricos. E, é claro, há uma análoga da hipótese de Riemann: todos os zeros não triviais da função zeta de Dedekind estão sobre a linha crítica. A expressão "hipótese generalizada de Riemann" inclui agora também esta conjectura.

Mesmo essa generalização não é o fim da história da função zeta. Ela inspirou a definição de funções análogas em várias outras áreas da matemática, desde álgebra abstrata até a teoria de sistemas dinâmicos. Em todas essas áreas há análogas da hipótese de Riemann de alcance ainda maior. Algumas delas chegaram a ser provadas verdadeiras. Em 1974, Pierre Deligne provou uma delas para variedades sobre campos finitos. Generalizações conhecidas como funções zeta de Selberg satisfazem uma

análoga da hipótese de Riemann. O mesmo vale para a função zeta de Goss. No entanto, existem outras generalizações, as funções zeta de Epstein, para as quais a análoga apropriada da hipótese de Riemann é falsa. Aqui infinitos zeros não triviais estão sobre a linha crítica, mas alguns não estão, como demonstrou Edward Titchmarsh em 1986. Por outro lado, essas funções zeta não possuem uma fórmula de produto de Euler, de modo que deixam de se parecer com as funções zeta de Riemann num aspecto que pode muito bem ser crucial.

A EVIDÂNCIA CIRCUNSTANCIAL em favor da veracidade da hipótese de Riemann – seja a original, sejam suas generalizações – é extensa. Muitas belas coisas seriam consequência da autenticidade dessa hipótese. Nenhuma delas jamais foi contestada: fazê-lo seria refutar a hipótese de Riemann, mas não se conhece nem prova nem refutação. Há uma sensação difundida de que a prova da hipótese de Riemann original abriria caminho para provar também suas generalizações. Na verdade, seria melhor atacar a hipótese generalizada de Riemann em toda a sua glória, explorando a riqueza de métodos atualmente disponíveis, e então deduzir a hipótese original como um caso especial.

Há também uma vasta quantidade de evidência experimental para a verdade da hipótese de Riemann – ou o que certamente parece ser uma vasta quantidade, até que alguém jogue água fria nessa alegação. Segundo Carl Ludwig Siegel, Riemann calculou os primeiros zeros da sua função zeta numericamente, mas não publicou os resultados: eles se localizam em

½ ± 14,135i ½ ± 21,022i ½ ± 25,011i

Os zeros não triviais sempre vêm em pares ± como estes. Escrevi aqui ½ em vez de 0,5 porque nesses casos a parte real é conhecida *exatamente*, explorando resultados gerais em análise complexa e propriedades conhecidas da função zeta. O mesmo vale para os cálculos computadorizados relatados abaixo. Eles não mostram apenas que os zeros estão muito próximos da linha crítica; na verdade estão nela.

Em 1903, Jorgen Gram mostrou numericamente que os primeiros dez (pares ±) de zeros jazem sobre a linha crítica. Em 1935, Titchmarsh havia aumentado o número para 195. Em 1936, Titchmarsh e Leslie Comrie provaram que os primeiros 1.041 pares de zeros estão sobre a linha crítica – e foi a última vez que alguém fez tais cálculos a mão. Alan Turing é mais

conhecido por seus esforços em tempo de guerra em Bletchley Park, onde ajudou a quebrar o código alemão Enigma, e pelo seu trabalho nos fundamentos da computação e da inteligência artificial. Mas Turing também se interessou pela teoria analítica dos números. Em 1953, descobriu um método mais eficiente para calcular zeros da função zeta, e usou um computador para deduzir que os primeiros 1.104 pares de zeros jazem sobre a linha crítica. Evidências de que todos os zeros até determinado limite estavam sobre a linha crítica foram se acumulando; o recorde atual, obtido por Yannick Saouter e Patrick Demichel, em 2004, é de 10 trilhões (10^{13}). Diversos matemáticos e cientistas da computação também verificaram outras gamas de zeros. Até a presente data, todo zero não trivial calculado jaz sobre a linha crítica.

Isso poderia parecer conclusivo, mas os matemáticos são ambivalentes em relação a esse tipo de evidência, e com razão. Números como 10 trilhões podem soar grandes, mas em teoria dos números o que geralmente importa é o logaritmo natural do número, que é proporcional à quantidade de dígitos. O logaritmo natural de 10 trilhões fica um pouco abaixo de 30. Na verdade, muitos problemas se articulam sobre o logaritmo do logaritmo, ou mesmo o logaritmo do logaritmo do logaritmo. Nesses termos, 10 trilhões é *ínfimo*, de modo que a evidência numérica até 10 trilhões mal carrega algum peso.

Há também alguma evidência geral analítica, que não está sujeita a essa objeção. Hardy e Littlewood provaram que infinitos zeros jazem sobre a linha crítica. Outros matemáticos mostraram, num sentido preciso, que quase todos os zeros estão muito próximos à linha crítica. Selberg provou que uma proporção não zero de zeros jaz sobre a linha crítica. Norman Levinson provou que essa proporção é de pelo menos um terço, cifra que agora foi melhorada para pelo menos 40%. Todos esses resultados sugerem que se a hipótese de Riemann for falsa, zeros que não estejam sobre a linha crítica são muito grandes, e muito raros. Infelizmente, a principal implicação é de que, se tais exceções existirem, encontrá-las será extraordinariamente difícil.

POR QUE SE INCOMODAR? Sem dúvida a evidência numérica deveria satisfazer qualquer pessoa sensata, não? Infelizmente, não. Não satisfaz os matemáticos, e nesse caso não estão apenas sendo pedantes: estão de fato agindo como pessoas sensatas. Em matemática geralmente, e sobretudo em

teoria dos números, aparentemente a evidência extensiva “experimental” muitas vezes tem um peso muito menor do que se possa imaginar.

Uma lição objetiva é fornecida pela conjectura de Pólya, enunciada em 1919 pelo matemático húngaro George Pólya. Ele sugeriu que pelo menos metade de todos os números inteiros até qualquer valor específico tem uma quantidade ímpar de fatores primos. Aqui fatores repetidos são contados separadamente, e começamos pelo 2. Por exemplo, até 20 a quantidade de fatores primos verifica-se conforme a Tabela 2, onde a coluna “percentagem” indica a percentagem de números até esse valor com quantidade ímpar de fatores primos.

Todas as percentagens na coluna final são superiores a 50%, e cálculos mais extensos tornam razoável conjecturar que isso é sempre verdade. Em 1919, sem computadores à disposição, experimentos não conseguiram encontrar nenhum número que refutasse tal conjectura. Mas em 1958 Brian Haselgrove usou a teoria analítica dos números para provar que a conjectura é falsa para algum número – menor que $1{,}845 \times 10^{361}$, para ser preciso. Uma vez os computadores em cena, Sherman Lehman mostrou que a conjectura é falsa para 906.180.359. Em 1980, Minoru Tanaka provou que o menor desses exemplos é 906.150.257. Desse modo, teria sido possível acumular evidência experimental em favor da conjectura para todos os números até quase 1 bilhão, ainda que ela seja falsa.

NÚMERO	FATORAÇÃO	QUANTOS PRIMOS?	PERCENTAGEM
2	2	1	100
3	3	1	100
4	2^2	2	66
5	5	1	75
6	2×3	2	60
7	7	1	66
8	2^3	3	71
9	3^2	2	62
10	2×5	2	55
11	11	1	60
12	$2^2 \times 3$	3	63
13	13	1	66
14	2×7	2	61
15	3×5	2	57
16	2^4	4	60
17	17	1	62
18	2×3^2	3	64
19	19	1	66
20	$2^2 \times 5$	3	65

TABELA 2 Percentagens de números, até um determinado valor, que têm uma quantidade ímpar de fatores primos.

Mesmo assim, é gostoso saber que o número 906.150.257 é especialmente interessante.

É claro que os computadores de hoje, se adequadamente programados, refutam essa conjectura em questão de segundos. Mas às vezes até mesmo computadores não adiantam. Um exemplo clássico é o número de Skewes, onde aparentemente imensas quantidades de evidência numérica a princípio sugeriam que uma famosa conjectura devia ser verdadeira, mas na verdade era falsa. Esse número gigantesco surgiu em um problema intimamente relacionado com a hipótese de Riemann: a aproximação de $\pi(x)$ por $\mathrm{Li}(x)$. Como acabamos de ver, o teorema dos números primos

afirma que a razão entre essas duas grandezas tende a 1 à medida que x torna-se grande. Cálculos numéricos parecem indicar algo mais forte: a razão é sempre menor que 1, ou seja, $\pi(x)$ é menor que $\mathrm{Li}(x)$. Em 2008, as computações numéricas de Tadej Kotnik mostraram que isso é verdade sempre que x for menor que 10^{14}. Em 2012, Douglas Stoll e Demichel haviam melhorado esse limite para 10^{18}, cifra obtida de maneira independente por Andry Kulsha. Resultados de Tomás Oliveira e Silva sugerem que pode ser aumentado para 10^{20}.

Isso pode soar definitivo. É mais consistente do que os melhores resultados numéricos que temos para a hipótese de Riemann. Mas em 1914 Littlewood provou que essa conjectura é falsa – incrivelmente falsa. À medida que x corre pelos números reais positivos, a diferença $\pi(x) - \mathrm{Li}(x)$ muda de sinal (de negativo para positivo, ou o inverso) com *infinita frequência*. Em particular, $\pi(x)$ é *maior* que $\mathrm{Li}(x)$ para valores suficientemente grandes de x. No entanto, a prova de Littlewood não dava indicação do tamanho desse valor.

Em 1933, um aluno seu, o matemático sul-africano Stanley Skewes, estimou qual deve ser o tamanho de x: nada mais que 10^10^10^34, onde ^ indica "elevado à potência". Esse número é tão gigantesco que se todos os dígitos fossem impressos em um livro – um livro bastante enfadonho, consistindo de um 1 seguido de intermináveis zeros – o universo não seria grande o suficiente para contê-lo, mesmo que cada dígito tivesse o tamanho de uma partícula subatômica. Mais ainda, Skewes teve de assumir a autenticidade da hipótese de Riemann para fazer sua prova funcionar. Em 1955, ele havia encontrado um meio de evitar a hipótese de Riemann, mas isso teve um preço: sua estimativa cresceu para 10^10^10^963.

Esses números são grandes demais até mesmo para o adjetivo "astronômico", porém pesquisa posterior os reduziu a algo que pode ser denominado cosmológico. Em 1966, Lehman substituiu os números de Skewes por $10^{1.165}$. Te Riele reduziu esse valor para 7×10^{370} em 1987, e em 2000 Carter Bays e Richard Hudson o reduziram a $1{,}39822 \times 10^{316}$. Kuok Fai Chow e Roger Plymen cortaram mais um pouco, e reduziram o número a $1{,}39801 \times 10^{316}$. Pode parecer uma melhora desprezível, mas é cerca de 2×10^{313} menor. Saouter e Demichel fizeram outra melhora para $1{,}3971667 \times 10^{316}$. Entretanto, em 1941, Aurel Wintner havia provado que uma pequena proporção, mas diferente de zero, de inteiros satisfaz $\pi(x) > \mathrm{Li}(x)$. Em 2011, Stoll e Demichel computaram os primeiros 200 bilhões de zeros da função zeta, o que dá controle sobre $\pi(x)$ quando x é algo até $10^{10.000.000.000.000}$, e

descobriram evidência de que se x for menor que $3{,}17 \times 10^{114}$ então $\pi(x)$ é menor que $\mathrm{Li}(x)$.[8] Logo, para este problema particular, a evidência pelo menos até 10^{18}, e muito possivelmente até 10^{114} ou mais, é completamente enganosa. Os volúveis deuses da teoria dos números estão rindo de uma boa piada à custa dos seres humanos.

NO DECORRER DOS ANOS, foram feitas muitas tentativas de provar ou refutar a hipótese de Riemann. O website de Matthew Watkins, "Proposed proofs of the Riemann hypothesis" (Provas propostas da hipótese de Riemann) lista por volta de cinquenta delas desde 2000.[9] Foram encontrados erros em muitas dessas tentativas, e nenhuma delas foi aceita como correta pelos peritos qualificados.

Um dos esforços mais amplamente publicados, em anos recentes, foi o de Louis de Branges, em 2002. Ele fez circular um longo manuscrito alegando deduzir a hipótese de Riemann aplicando um ramo da análise que lidava com operadores em espaços de infinitas dimensões, conhecido como análise funcional. Havia motivos para levar De Branges a sério. Anteriormente ele fizera circular uma prova da conjectura de Bieberbach sobre expansão serial de funções complexas. Embora sua prova original contivesse erros, acabou ficando estabelecido que a ideia subjacente funcionava. Todavia, agora parece haver boas razões para pensar que o método proposto por De Branges para provar a hipótese de Riemann não tem chance de dar certo. Alguns obstáculos aparentemente fatais foram assinalados por Brian Conrey e Xian-Jin Li.[10]

Talvez a maior esperança de uma prova venha de formas novas ou radicalmente diferentes de se pensar a respeito do problema. Como vimos com frequência, impulsos significativos em grandes problemas muitas vezes surgem quando alguém os relaciona a alguma área da matemática totalmente diferente. O último teorema de Fermat é um exemplo claro: uma vez reinterpretado como questão sobre curvas elípticas, o progresso foi rápido.

A tática de De Branges agora parece questionável, mas sua abordagem é estrategicamente sólida. Tem raízes numa sugestão verbal feita em torno de 1912 por David Hilbert e, de modo independente, por George Pólya. O físico Edmund Landau pediu a Pólya uma razão física para que a hipótese de Riemann devesse ser considerada verdadeira. Pólya relatou em 1982 que se deparara com uma resposta: os zeros da função zeta devem estar

relacionados com os autovalores do chamado operador autoadjunto. Esses são números característicos associados com tipos especiais de transformação. Em física quântica, uma das aplicações importantes, esses números determinam os níveis de energia do referido sistema, e um teorema padrão simples afirma que os autovalores desse tipo especial de operador são sempre reais. Como vimos, a hipótese de Riemann pode ser reformulada como a afirmação de que todos os zeros da função xi são reais. Se algum operador autoadjunto tivesse autovalores que fossem os mesmos que os zeros da função xi, a hipótese de Riemann seria uma consequência fácil. Pólya não publicou essa ideia – não conseguiu registrar por escrito tal operador, e quando alguém conseguisse, seria o paraíso na Terra. Mas em 1950 Selberg provou sua "fórmula de traçado", que relaciona a geometria de uma superfície com os autovalores de um operador associado. Isso fez a ideia parecer um pouco mais plausível.

Em 1972, Hugh Montgomery estava visitando o Instituto de Estudos Avançados em Princeton. Ele havia notado algumas características estatísticas surpreendentes nos zeros não triviais da função zeta. Mencionou-as ao físico Freeman Dyson, que imediatamente identificou uma semelhança com características estatísticas de matrizes hermitianas aleatórias, outro tipo especial de operador usado para descrever sistemas quânticos tais como núcleos atômicos. Em 1999, Alain Connes surgiu com uma fórmula de traçado, similar à de Selberg, cuja validade implicaria a autenticidade da hipótese generalizada de Riemann. Também em 1999, os físicos Michael Berry e Jon Keating sugeriram que o operador exigido poderia surgir quantizando-se um conceito bem conhecido da física clássica, relacionado com a quantidade de movimento. A resultante conjectura de Berry pode ser vista como uma versão mais específica da conjectura de Hilbert-Pólya.

Essas ideias, relacionando a hipótese de Riemann com áreas centrais em física matemática, são extraordinárias. Mostram que o progresso pode finalmente vir de áreas da matemática aparentemente não relacionadas, e levanta a esperança de que a hipótese de Riemann possa um dia ser solucionada. Contudo, tais ideias ainda não levaram a nenhum avanço definitivo que nos encoraje a pensar que a solução está logo ali dobrando a esquina. A hipótese de Riemann continua sendo um dos enigmas mais desconcertantes e irritantes de toda a matemática.

HOJE HÁ UMA NOVA RAZÃO para tentar provar a hipótese de Riemann: um prêmio substancial.

Não existe prêmio Nobel de matemática. O mais distinto prêmio nessa disciplina é a medalha Fields, mais apropriadamente chamada Medalha Internacional para Descobertas Excepcionais em Matemática. Ela recebe o nome do matemático canadense John Fields, que doou o prêmio em seu testamento. A cada quatro anos, no Congresso Internacional de Matemáticos, até quatro dos mais proeminentes matemáticos jovens do mundo (com menos de quarenta anos) recebem uma medalha de ouro e um prêmio em dinheiro, atualmente 15 mil dólares. No que diz respeito à matemática, a medalha Fields é equivalente em prestígio ao prêmio Nobel.

Muitos matemáticos consideram uma boa coisa a falta de um Nobel nessa matéria. Um prêmio Nobel atualmente vale pouco mais de 1 milhão de dólares, quantia que poderia com facilidade distorcer objetivos de pesquisas e levar a discussões sobre primazia. No entanto, a ausência de um prêmio matemático de primeira grandeza pode também ter distorcido a percepção pública do valor e da utilidade da matemática. É fácil imaginar que se ninguém está disposto a pagar, não deve valer muita coisa.

Há pouco tempo, dois novos prêmios matemáticos de grande prestígio passaram a existir. Um é o Abel, concedido anualmente pela Academia Norueguesa de Ciências e Letras, e batizado em homenagem ao grande matemático norueguês Niels Henrik Abel. O outro consiste em sete prêmios do milênio do Instituto Clay de Matemática. Essa instituição foi fundada por Landon Clay e sua mulher, Lavinia. Landon Clay é um empresário norte-americano ativo em fundos mútuos, com amor e respeito pela matemática. Em 1999, estabeleceu uma nova fundação para matemática em Cambridge, Massachusetts, que organiza encontros, concede verbas para pesquisas, prepara conferências públicas e administra um prêmio anual de pesquisa.

Em 2000, sir Michael Atiyah e John Tate, importantes matemáticos na Grã-Bretanha e nos Estados Unidos, anunciaram que o Instituto Clay de Matemática havia criado um novo prêmio, destinado a incentivar a solução de sete dos mais importantes problemas em aberto da matemática. Passariam a ser conhecidos como os problemas do milênio, e uma solução apropriadamente publicada e referendada de qualquer um deles renderia 1 milhão de dólares. Juntos, esses problemas chamavam atenção para algumas das questões centrais não respondidas na matemática, cuidadosamente selecionadas pelos mais importantes matemáticos do mundo. O prêmio substancial deixa um ponto muito claro ao público: a

matemática é valiosa. Todos os envolvidos estão cientes de que seu valor intelectual pode ser mais profundo do que meramente o dinheiro, mas um prêmio em dinheiro vivo ajuda a cabeça a se concentrar. O mais conhecido prêmio do milênio é a hipótese de Riemann. É a única questão que aparece tanto na lista de Hilbert de 1900 como na lista dos problemas do milênio. Os outros seis problemas do milênio são discutidos nos capítulos 10 a 15. Os matemáticos não são especialmente obcecados por prêmios, e trabalhariam na hipótese de Riemann com ou sem ele. Uma ideia nova e promissora seria toda a motivação necessária. Vale a pena lembrar que conjecturas, por mais que sejam reverenciadas pela sua antiguidade, podem não ser verdadeiras. Hoje, a maioria dos matemáticos parece pensar que uma prova para a hipótese de Riemann pode vir a ser encontrada. Alguns, porém, acreditam que ela pode ser falsa: em algum lugar lá no meio da imensidão bravia dos números muito grandes pode estar espreitando um zero que não esteja sobre a linha crítica. Se tal "contraexemplo" existir, é provável que seja muito, muito grande.

Entretanto, opiniões contam pouco nas fronteiras da matemática. A intuição dos conhecedores com frequência é de fato muito boa, mas tem havido inúmeras ocasiões em que estava errada. A sabedoria formal pode ser ao mesmo tempo sábia e convencional, sem ser verdadeira. Littlewood, um dos grandes entendidos em análise complexa, foi inequívoco: em 1962 disse que tinha certeza de que a hipótese de Riemann era falsa, acrescentando que não havia razão imaginável para ser verdadeira. Quem está certo? Podemos apenas esperar para ver.

10. Qual é o formato de uma esfera?

A conjectura de Poincaré

HENRI POINCARÉ foi um dos mais importantes matemáticos do fim do século XIX, um pouquinho excêntrico, porém um operador astuto. Tornou-se membro do Bureau des Longitudes da França, cuja tarefa era aperfeiçoar a navegação, medidas de tempo e medições da Terra e dos planetas. Essa indicação o levou a propor o estabelecimento de zonas de tempo internacionais; e também o inspirou a pensar sobre a física do tempo, antecipando algumas das descobertas de Einstein na relatividade especial. Poincaré deixou sua marca por todo o panorama matemático, desde a teoria dos números até a física matemática.

Foi sobretudo um dos fundadores da topologia, a matemática das transformações contínuas. Aqui, em 1904, deparou-se com uma questão aparentemente simples, tendo percebido mais tarde que admitira tacitamente a resposta em trabalho anterior, mas não conseguiu encontrar uma prova. "Esta pergunta nos levaria para muito longe", escreveu, o que na verdade ocultava a coisa real: a pergunta não o estava levando *a lugar algum*. Embora tivesse formulado o problema em termos de pergunta, ele se tornou conhecido como conjectura de Poincaré porque todo mundo esperava que a resposta fosse "sim". É outro dos sete problemas do milênio do prêmio Clay, e merecidamente, porque acabou se revelando um dos problemas mais desconcertantes em toda a topologia. A pergunta de Poincaré foi finalmente respondida em 2002 por um jovem russo, Grigori Perelman. A solução introduzia toda uma série de novas ideias e métodos, tanto que a comunidade matemática levou alguns anos para digerir a prova e aceitar que estava correta.

Pelo seu sucesso, Perelman foi laureado com a medalha Fields, o mais prestigioso prêmio matemático, mas declinou. Não queria publicidade. Foi-lhe oferecido o prêmio Clay de 1 milhão de dólares por provar a conjectura de Poincaré, e ele recusou. Tampouco queria dinheiro. O que Perelman desejava era que seu trabalho fosse aceito pela comunidade matemática. Por fim acabou sendo, mas infelizmente, por razões sensatas, isso levou algum tempo. Além disso, é irreal esperar aceitação sem publicidade ou

oferta de prêmios. Mas essas consequências inevitáveis do sucesso não se encaixavam na natureza às vezes reclusa de Perelman.

ENCONTRAMOS A TOPOLOGIA em conexão com o teorema das quatro cores, e recorremos ao clichê "geometria da folha de borracha". A geometria de Euclides lida com linhas retas, círculos, comprimentos e ângulos. Ela tem lugar no plano, ou num espaço de três dimensões quando se torna mais avançada. Um plano é como uma folha de papel infinita, e compartilha uma característica básica do papel: não estica, não encolhe e não se curva. Pode-se enrolar o papel de maneira a formar um tubo, e ele pode esticar ou encolher um pouquinho, sobretudo se você derramar café em cima. Mas você não consegue embrulhar uma esfera com uma folha de papel sem criar dobras. Matematicamente, o plano euclidiano é rígido. Na geometria de Euclides, dois objetos – triângulos, quadrados, círculos – são iguais se pudermos transformar um no outro mediante um deslocamento rígido. E "rígido" quer dizer que as distâncias não se alteram.

E se você usar uma folha elástica em vez de papel? Ela estica, ela se curva e, com um pouco de esforço, pode ser comprimida. Comprimentos e ângulos não têm significado fixo numa folha elástica. De fato, se ela for elástica *o bastante*, tampouco têm sentido triângulos, quadrados ou círculos. Você pode deformar um triângulo numa folha de borracha de modo a lhe dar um vértice a mais. Pode até transformá-lo em um círculo (Figura 36). Quaisquer que sejam os conceitos da geometria da folha de borracha, eles não incluem os euclidianos tradicionais.

FIGURA 36 Deformação topológica de um triângulo em um círculo.

Poderia parecer que a geometria numa folha de borracha é tão flexível que nada teria sentido fixo, e nesse caso pouca coisa substancial poderia ser provada. Não é assim. Desenhe um triângulo e coloque um ponto dentro dele. Se você esticar e deformar a folha até o triângulo virar um círculo, uma característica do seu desenho não se altera: o ponto permanece do lado de dentro. Certo, agora ele está dentro de um círculo, não de um

triângulo, mas não está *fora*. Para levar o ponto para fora é preciso rasgar a folha. Isso quebra as regras desse jogo específico.

Há outra característica que também sobrevive à distorção. Um triângulo é uma curva fechada simples. É uma linha que se junta a si mesma de modo que não haja pontas livres, e essa linha não cruza sobre si mesma. O número oito é uma curva fechada, mas não é simples – ele cruza sobre si mesmo. Quando você deforma a folha de borracha, o triângulo pode mudar de formato, mas sempre permanece uma curva fechada simples. Não há como transformá-lo em um oito, por exemplo, sem rasgar a folha.

Em topologia tridimensional, a totalidade do espaço torna-se elástica. Não como um bloco de borracha, que volta a assumir seu formato original se você o largar, mas como um gel que pode ser deformado sem qualquer resistência. Um espaço topológico é infinitamente deformável. Você pode pegar uma região do tamanho de um grão de arroz e ampliá-la até o tamanho do Sol. Pode puxar tentáculos até que a região tenha o formato de um polvo. A única coisa que você não tem permissão para fazer é introduzir qualquer tipo de descontinuidade. Não pode rasgar o espaço nem executar qualquer tipo de distorção que provoque fendas entre dois pontos próximos.

Que características de uma forma no espaço sobrevivem a todas as deformações contínuas? Comprimento não, nem área, nem volume. Mas se nessa forma de espaço existirem nós, estes sobreviverão. Se você der um nó numa curva e juntar as pontas para criar um laço, então o nó não consegue se desfazer. Por mais que você deforme o espaço, a curva continua com o nó. Assim, estamos trabalhando com um novo tipo de geometria na qual os conceitos importantes e significativos parecem bastante vagos: “dentro”, “fechado”, “simples”, “com nó”. Essa nova geometria tem um nome respeitável: topologia. Ela pode parecer meio esotérica, talvez até absurda, mas tornou-se uma das principais áreas da matemática do século XX, e permanece igualmente vital no século XXI. E uma das principais pessoas a quem devemos agradecer por isso é Poincaré.

A HISTÓRIA DA TOPOLOGIA começou a decolar quase um século antes de Poincaré, em 1813. Simon Antoine Jean Lhuilier, um matemático suíço, não foi exatamente um dos luminares da matemática de seu tempo, embora tenha rejeitado uma grande soma em dinheiro que um parente lhe prometera caso entrasse para a Igreja. Lhuilier preferiu uma carreira em

matemática. Especializou-se em um nicho da matemática pouco visível: o teorema de Euler para poliedros. No Capítulo 4 deparamo-nos com esse resultado curioso, aparentemente isolado: se um poliedro tem F faces, V vértices e A arestas, então $F - A + V = 2$. Lhuilier passou grande parte da sua carreira investigando variantes dessa fórmula, e em retrospecto deu um passo vital na direção da topologia quando descobriu que a fórmula de Euler às vezes está errada. Sua validade depende do formato qualitativo do poliedro.

A fórmula é correta para poliedros sem furos, que podem ser desenhados sobre a superfície de uma esfera ou deformados continuamente numa forma desse tipo. Mas quando o poliedro tem furos, a fórmula falha. Uma moldura de quadro feita de madeira com uma seção transversal retangular, por exemplo, tem dezesseis faces, 32 arestas e dezesseis vértices; aqui $F - A + V = 0$. Lhuilier modificou a fórmula de Euler para cobrir esses poliedros mais exóticos: se houver g furos, então $F - A + V = 2 - 2g$. Essa foi a primeira descoberta de um invariante topológico significativo: uma grandeza associada a um espaço, que não varia quando o espaço é deformado continuamente. O invariante de Lhuilier fornece um modo rigoroso de contar quantos furos uma superfície tem, sem a necessidade de definir "furo". Isso é útil, porque o conceito de furo é traiçoeiro. Um furo não é parte da superfície, e tampouco a região fora da superfície. Ele parece ser uma característica de como a superfície se assenta no espaço circundante. Mas a descoberta de Lhuilier mostra que o que interpretamos como *número* de furos é uma característica intrínseca, independente de qualquer espaço circundante. Não é necessário definir furos e então contá-los; na verdade, é melhor não.

Depois de Lhuilier, a figura-chave seguinte na pré-história da topologia é Gauss. Ele encontrou diversos outros invariantes topológicos ao trabalhar em várias áreas centrais da matemática. Seu trabalho em análise complexa, sobretudo a prova de que toda equação polinomial tem pelo menos uma solução em números complexos, o levou a considerar o número de voltas de uma curva no plano: quantas vezes a curva dá a volta em torno de um dado ponto. Problemas em eletricidade e magnetismo levaram ao número de ligação de duas curvas fechadas: quantas vezes uma gira em torno da outra. Estes e outros exemplos levaram Gauss a perguntar-se se poderia existir algum ramo ainda não descoberto da matemática que fornecesse um meio sistemático para compreender características qualitativas de figuras geométricas. Ele não publicou coisa alguma a respeito do tópico, mas o

mencionou em cartas e manuscritos.

Gauss também passou a ideia adiante para seu aluno Johann Listing e seu assistente August Möbius. Já mencionei a faixa de Möbius, uma superfície com um lado só e também uma aresta só, que ele publicou em 1865 e pode ser encontrada na Figura 9 do Capítulo 4. Möbius ressaltou que "ter um lado só", ainda que intuitivamente claro, é difícil de precisar, e propôs uma propriedade relacionada que podia ser definida com completo rigor. Essa propriedade era a orientabilidade. Uma superfície é orientável se você puder cobri-la com uma rede de triângulos, com setas circundando cada triângulo, de modo que sempre que dois triângulos tiverem uma aresta comum as setas apontem em sentidos opostos. É o que acontece se você desenhar uma rede num plano e fizer as setas apontarem no sentido horário, por exemplo. Numa faixa de Möbius, não existe uma rede dessas.

A primeira publicação de Listing em topologia veio antes, em 1847. Seu título era *Vorstudien zur Topologie* (Estudos preliminares para a topologia), e foi o primeiro texto a empregar essa palavra. Ele vinha usando o termo de maneira informal por cerca de uma década. Outro termo usado na época era a expressão latina *analysis situ*, "análise de posição", mas esta acabou caindo em desuso. O livro de Listing contém pouca coisa de grande importância, mas estabelece, sim, uma noção básica: cobrir uma superfície com uma rede de triângulos. Em 1861, quatro anos antes de Möbius, ele descreveu a faixa de Möbius e estudou conectividade: se um espaço pode ser dividido em duas ou mais partes desconectadas. Elaborando o trabalho de Listing, alguns matemáticos, entre eles Walther von Dyck, reuniram uma classificação topológica completa de superfícies, assumindo-as fechadas (sem arestas) e compactas (de extensão finita). A resposta é que toda superfície orientável é topologicamente equivalente a uma esfera, à qual foi atado um número finito g de alças (Figura 11, centro e direita), no Capítulo 4. O número g é chamado genus da superfície, e é o que determina o invariante de Lhuilier. Se $g = 0$ temos a esfera, e se $g > 0$ obtemos um toro com g furos. Uma sequência similar de superfícies, começando com a superfície não orientável mais simples, o plano projetivo, classifica todas as superfícies não orientáveis. O método foi estendido de modo a permitir também superfícies com bordas. Cada borda é um laço fechado, e a única informação necessária adicional é quantos desses laços ocorrem.

A CONJECTURA DE POINCARÉ fará mais sentido se primeiro dermos uma olhada

em uma das técnicas básicas empregadas em classificar superfícies. Anteriormente, descrevi a topologia em termos de deformar uma figura feita de borracha ou gel, e enfatizei a necessidade de usar transformações *contínuas*. Por ironia, uma das técnicas centrais em topologia envolve o que à primeira vista pode parecer uma transformação descontínua: cortar uma figura em pedaços. No entanto, a continuidade é restaurada por uma série de regras, descrevendo qual pedaço juntar com qual, e de que maneira. Um exemplo é o modo como definimos um toro identificando lados opostos de um quadrado (Figura 12 do Capítulo 4).

Identificar pontos que parecem ser distintos permite-nos representar espaços topológicos complicados usando ingredientes simples. Um quadrado é um quadrado e pronto, mas um quadrado com regras de identificação pode ser um toro, uma garrafa de Klein, um cilindro, uma faixa de Möbius ou um plano projetivo, dependendo da escolha das regras (Figura 37). Assim, quando expliquei uma transformação contínua em termos de esticar e curvar uma folha de borracha, pedi mais do que o estritamente necessário. Temos também permissão para cortar a folha num estágio intermediário, com a condição de que acabemos fazendo uma das duas coisas: ou voltamos a juntar as bordas exatamente como estavam no começo ou especificamos regras que tenham o mesmo efeito. No que diz respeito à topologia, *enunciar* uma regra para colar as bordas é o mesmo que efetivamente implantar a regra. Contanto que você não esqueça a regra no que vier a fazer mais tarde.

O método clássico para classificar superfícies começa por desenhar uma rede de triângulos sobre a superfície. Aí cortamos lados suficientes dos triângulos para conseguir dobrá-los de modo a formar um polígono. Regras de colagem, derivadas de como fazemos os cortes, especificam então como identificar as várias bordas do polígono, reconstituindo a superfície original. Nesse ponto, toda a topologia interessante está implícita nas regras de colagem. A classificação é provada manipulando-se as regras algebricamente e transformando-as em regras que definem um toro com g furos ou uma das superfícies não orientáveis análogas. A topologia moderna tem outros meios para obter o mesmo resultado, mas muitas vezes usa essa espécie de construção “recortar e colar”. O método se generaliza sem dificuldade para espaços de qualquer dimensão, mas é restrito demais para levar a uma classificação de espaços topológicos de dimensões superiores sem assistência adicional.

FIGURA 37 Cinco espaços topológicos diferentes obtidos fazendo coincidir lados opostos de um quadrado de várias maneiras.

POR VOLTA DE 1900, Poincaré estava desenvolvendo o trabalho inicial na topologia de superfícies por meio de uma técnica mais genérica, que se aplicava a espaços com qualquer número de dimensões. A principal investida de sua pesquisa era descobrir invariantes topológicos: números ou fórmulas algébricas associadas a espaços, que se conservam imutáveis quando o espaço é deformado continuamente. Se dois espaços têm invariantes diferentes, não podem ser deformados em outro, então são topologicamente distintos.

Poincaré começou a partir da generalização feita pelo matemático italiano Enrico Betti do invariante topológico de Lhuilier, $F - A + V$, agora muito injustamente conhecido como característica de Euler, para espaços de dimensões superiores, obtido em 1870. Betti havia notado que a maior quantidade de curvas fechadas que podem ser desenhadas numa superfície de genus g, sem dividi-la em partes desconexas, é $g - 1$. Essa é outra maneira de caracterizar topologicamente a superfície. Ele generalizou essa ideia em "números de conectividade" de qualquer dimensão, que Poincaré chamou de números de Betti, um termo usado ainda hoje. O número de Betti k-dimensional conta a quantidade de furos k-dimensionais no espaço.

Poincaré desenvolveu os números de conectividade de Betti em um invariante mais sensível chamado homologia, que tem muito mais estrutura algébrica. Discutiremos homologia mais detalhadamente no Capítulo 15. Basta dizer que ela examina coleções de "faces" multidimensionais nesse tipo de rede, e pergunta quais delas formam a fronteira de um disco topológico. Um disco não tem furos, ao contrário de um toro, de modo que podemos ter certeza de que dentro de qualquer coleção de faces que constitua uma fronteira não há furos. Ao contrário, podemos detectar furos contrastando coleções de faces que não formam fronteiras com coleções que formam. Dessa maneira, podemos construir uma série de invariantes de um espaço, conhecidos como grupos de

homologia. “Grupo”, aqui, é um termo da álgebra abstrata; significa que dois objetos quaisquer no grupo podem ser combinados para gerar outra coisa no mesmo grupo, de uma maneira sujeita a diversas regras algébricas bacanas. Falarei adiante um pouco mais a esse respeito, quando precisarmos dessa ideia. Existe um grupo desses para cada dimensão de 0 a *n*, e para cada espaço obtemos uma série de invariantes topológicos, com todo tipo de propriedades algébricas fascinantes.

Listing havia classificado todas as superfícies topológicas – espaços de dimensão 2. O passo óbvio seguinte era olhar a dimensão 3. E o espaço mais simples para começar era uma esfera. Na linguagem corrente a palavra “esfera” tem dois significados diferentes: pode ser uma bola redonda sólida ou apenas a superfície da bola. Quando trabalhamos em topologia de superfícies, esse termo é sempre interpretado pelo segundo significado: a superfície infinitamente fina de uma bola. Além disso, o interior da esfera não é considerado como fazendo parte dela: é apenas uma consequência da maneira habitual como mergulhamos uma superfície esférica no espaço. Intrinsecamente, tudo que temos é uma superfície, equivalente, de maneira topológica, à superfície de uma bola. Você pode pensar na esfera como uma bola vazia com uma crosta infinitamente delgada.

A análoga 3-dimensional “correta” de uma esfera, chamada 3-esfera, *não* é uma bola sólida. Uma esfera não tem fronteira, e tampouco sua análoga 3-dimensional deve ter. O modo mais simples de definir uma 3-esfera é imitar a geometria de coordenadas de uma esfera comum. Isso leva a um espaço um pouco traiçoeiro de ser visualizado: não posso mostrar um modelo em três dimensões porque a 3-esfera – mesmo tendo apenas três dimensões – não está mergulhada no espaço 3-dimensional comum. Em vez disso, está mergulhada num espaço 4-dimensional.

A esfera unitária comum, no espaço 3-dimensional, consiste em todos os pontos distantes uma unidade de um ponto específico: o centro. De forma análoga, a 3-esfera unitária num espaço 4-dimensional consiste em todos os pontos que estão a uma distância unitária do centro. Em coordenadas podemos escrever uma fórmula para isso utilizando uma generalização do teorema de Pitágoras para definir a distância.[1] De modo geral, uma 3-esfera é *qualquer* espaço topologicamente equivalente à 3-esfera unitária, assim como todos os tipos de versões granulosas de uma 2-esfera unitária são 2-esferas topológicas, e, obviamente, o mesmo vale para dimensões superiores.

Se você não está satisfeito com isso e quer uma imagem mais geométrica, experimente esta aqui. Uma 3-esfera pode ser representada como uma bola sólida cuja superfície inteira é identificada com um único ponto. Esse é outro exemplo de uma regra de colagem, e nesse caso é análoga à maneira de transformar um disco circular numa 2-esfera. Se você passar uma corda ao redor da borda de um disco de pano, e apertar bastante, como se estivesse fechando um saco, o resultado é topologicamente o mesmo que uma 2-esfera. Agora realize uma operação análoga em uma bola sólida, mas, como sempre, não tente visualizar o resultado: apenas pense em uma bola sólida e implante as regras de colagem conceitualmente.

Em todo caso, Poincaré estava muito interessado na 3-esfera, porque era presumivelmente o espaço topológico 3-dimensional mais simples sem fronteira e de extensão finita. Em 1900, publicou um artigo no qual afirmava que grupos de homologia eram um invariante poderoso o suficiente para caracterizar topologicamente a 3-esfera. De maneira específica, se um espaço topológico 3-dimensional tem os mesmos grupos de homologia que uma 3-esfera, então é topologicamente equivalente a (pode ser continuamente deformado em) uma 3-esfera. Em 1904, porém, havia descoberto que a afirmação estava errada. Há pelo menos um espaço 3-dimensional que não é uma 3-esfera, mas possui os mesmos grupos de homologia que uma 3-esfera. O espaço foi um triunfo para a filosofia das regras de colagem, e a prova de que não era uma 3-esfera envolvia a criação de um novo invariante, necessariamente mais poderoso que a homologia.

Primeiro, o espaço. Ele é chamado espaço dodecaédrico de Poincaré porque uma construção moderna usa o sólido dodecaedro. Poincaré não estava ciente dessa relação com o dodecaedro: colou dois toros sólidos entre si de maneira muito obscura. A interpretação do dodecaedro foi publicada em 1933, cerca de 21 anos depois da morte de Poincaré, por Herbert Seifert e Constantin Weber, e é muito mais fácil de compreender. A analogia que se deve ter em mente é a construção de um toro colando lados opostos de um quadrado. Como sempre, não se tentar *fazer* a colagem; simplesmente nos lembramos de que pontos correspondentes são considerados o mesmo ponto. Agora fazemos a mesma coisa, mas usando faces opostas de um dodecaedro (Figura 38).

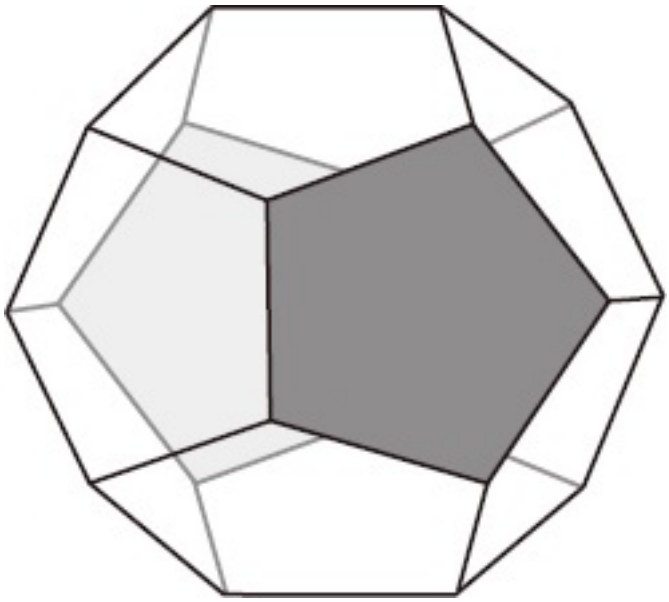


FIGURA 38 Para fazer o espaço dodecaédrico de Poincaré, pegue um dodecaedro e cole todos os pares de faces opostas (como o par sombreado), com uma torção para fazer com que se encaixem.

Os pitagóricos já conheciam os dodecaedros 2.500 anos atrás. A fronteira de um dodecaedro consiste em doze pentágonos regulares, unidos para formar uma gaiola aproximadamente esférica, com três pentágonos juntando-se em cada vértice. Agora cole cada face do dodecaedro com a face oposta… Só que há uma torção. Literalmente. Cada face precisa ser girada de um ângulo conveniente antes de ser grudada à face oposta. O ângulo é o menor capaz de alinhar as faces correspondentes, ou seja, 36 graus. Você pode pensar nessa regra como uma versão elaborada da regra da faixa de Möbius: torça uma borda a 180 graus e então cole-a com a borda oposta.

O espaço é esse. Agora vamos dar uma olhada no invariante. Não estou simplesmente divagando: precisamos de tudo isso para entender a conjectura de Poincaré.

POINCARÉ CHAMOU SEU novo invariante de grupo fundamental. Hoje ainda usamos esse nome, mas também nos referimos a ele como o (primeiro) grupo de homotopia. Homotopia é uma construção geométrica que pode ser executada inteiramente dentro do espaço, e fornece informação sobre o tipo topológico desse espaço. Ela o faz usando uma estrutura algébrica abstrata conhecida como grupo. Um grupo é uma coleção de objetos matemáticos, onde dois quaisquer podem ser combinados para dar outro objeto do grupo. Essa lei de combinação – frequentemente chamada multiplicação ou adição, mesmo quando não são as operações aritméticas

usuais com esses nomes – deve satisfazer algumas condições simples e naturais. Se chamarmos a operação de adição, as principais condições são:

- O grupo contém um elemento que se comporta como zero: se você somar a ele qualquer coisa do grupo, obtém aquela mesma coisa como resultado.
- Todo membro do grupo tem um negativo no grupo: some os dois e você obtém zero.
- Se você somar três membros do grupo, não importa quais deles você some antes. Ou seja:
  $(a + b) + c = a + (b + c)$.
  Essa é chamada de propriedade associativa.

A lei algébrica que *não* é imposta (embora às vezes também seja verdadeira) é a propriedade comutativa: $a + b = b + a$.[2]

O grupo fundamental de Poincaré é uma espécie de esqueleto simplificado do espaço. É um invariante topológico: espaços topologicamente equivalentes têm o mesmo grupo fundamental. Para obter uma visão prática, e possivelmente reconstituir parte do raciocínio de Poincaré, vamos ver como ele funciona para um círculo, apoderando-nos de uma imagem que remonta aos tempos de Gauss. Imagine uma formiga cujo universo inteiro seja o círculo. Como ela pode descobrir qual é o formato desse universo? Ela consegue distinguir um círculo de, digamos, uma reta? Tenha em mente o tempo todo que a formiga não pode sair do seu universo, olhar para ele e ver que é circular. Tudo que ela pode fazer é vagar dando voltas dentro do seu universo, qualquer que seja ele. Em particular, a formiga não perceberá que seu universo é curvo, porque sua versão de um raio de luz também está confinada a um círculo. Por favor, ignore aspectos práticos como os objetos terem de passar um pelo outro – será uma analogia bastante livre.

A formiga pode descobrir o formato do seu universo de diversas maneiras. Vou me concentrar em um método que se generaliza para qualquer espaço topológico. Para os propósitos desta discussão, a formiga é um ponto. E mora numa parada de ônibus, que também é um ponto. Todo dia ela sai de casa, pega o ônibus (que, naturalmente, é um ponto) e termina voltando para casa. A viagem mais direta é no ônibus número 0, que simplesmente fica estacionado na parada e não vai a lugar algum. Para

uma excursão mais interessante, a formiga pega o ônibus número 1, que dá a volta no universo exatamente uma vez em sentido anti-horário e para assim que chega de volta em casa. O ônibus número 2 dá duas voltas, o ônibus 3 dá três voltas, e assim por diante: um ônibus anti-horário para cada inteiro positivo. Há também ônibus negativos, que viajam no sentido oposto. O ônibus número −1 dá uma volta no sentido horário, o −2, duas voltas no sentido horário, e assim por diante.

A formiga logo percebe que duas viagens sucessivas no ônibus 1 são essencialmente a mesma coisa que uma única viagem no ônibus 2, e três viagens no 1 são o mesmo que uma única viagem no 3. De maneira semelhante, uma viagem no ônibus 5 seguida de uma viagem no 8 é essencialmente a mesma coisa que uma viagem no ônibus 13. Na verdade, dados dois números positivos quaisquer, uma viagem no ônibus com o primeiro número, seguida de uma viagem no ônibus com o segundo número, equivale a uma viagem no ônibus cujo número é sua soma.

O passo seguinte é mais sutil. A mesma relação *quase* vale para ônibus cujos números são negativos ou zero. Uma viagem no número 0, seguida de uma viagem no número 1, é muito similar a uma viagem no ônibus 1. No entanto, há uma leve diferença. Na viagem 0 + 1, o ônibus 0 espera por um tempo no começo, o que não acontece numa viagem só no ônibus 1. Assim, introduzimos uma noção com o proibitivo nome de homotopia (“mesmo lugar”, em grego). Duas voltas são homotópicas se uma puder ser continuamente deformada na outra. Se permitirmos que os itinerários de ônibus sejam alterados por homotopias, podemos gradualmente encolher o tempo até o período estacionário sumir. Agora, a diferença entre a viagem 0 + 1 e a viagem 1 desapareceu, então “quanto à homotopia” o resultado é simplesmente uma viagem no ônibus número 1. Ou seja, a equação com os números de ônibus 0 + 1 = 1 permanece válida – não para viagens, mas para classes de homotopia de viagens.

Que tal uma viagem no ônibus 1 seguida de uma viagem no −1? Nós igualaríamos isso a uma viagem no ônibus 0, mas não é. Ela percorre todo o caminho no sentido anti-horário, e depois volta percorrendo tudo no sentido horário. Fica bem claro que isso é diferente de passar a viagem inteira sentado dentro do ônibus 0 parado no ponto. Logo, 1 + (−1), isto é, 1 − 1, não é igual a 0. Porém, mais uma vez, a homotopia vem nos salvar: a combinação dos ônibus 1 e −1 é homotópica para a mesma viagem global do ônibus 0. Para ver por quê, suponha que a formiga siga de carro a rota combinada dos ônibus 1 e −1, mas um pouquinho antes de chegar ao ponto

de ônibus ela inverta o sentido e volte para casa. Essa viagem é bem próxima à viagem dupla de ônibus: só perde uma fração ínfima da viagem. Logo, a dupla viagem original de ônibus "encolheu", continuamente, para uma viagem de carro ligeiramente menor. Agora a formiga pode encurtar a viagem novamente, dando meia-volta um pouco antes. E pode continuar encolhendo a viagem, invertendo aos poucos o sentido cada vez mais cedo, até que, por fim, tudo que ela faz é ficar sentada num carro estacionado no ponto de ônibus, sem ir a lugar algum. Esse processo de encolhimento também é uma homotopia, e mostra que uma viagem no 1 seguida de uma viagem no −1 é homotópica a uma viagem no ônibus número 0. Ou seja, 1 + (−1) = 0 para classes de homotopia de viagens.

Agora é um caminho direto, para um algebrista, provar que uma viagem em qualquer ônibus, seguida de uma viagem num segundo ônibus, é homotópica ao ônibus que obtém somando-se os números dos dois ônibus. Isso vale para ônibus positivos, e o ônibus zero. Assim, se somarmos viagens de ônibus umas às outras – bem, classes de homotopia de viagens de ônibus – obtemos um grupo. Na verdade, é um grupo muito familiar. Seus elementos são os inteiros (números de ônibus) e sua operação é a adição. Seu símbolo convencional é $\mathbb{Z}$, da palavra alemã Zahl ("inteiro").

Um pouco mais de trabalho duro prova que em um universo circular *qualquer* viagem de carro dando a volta – ainda que envolva montes de recuos, inversão de sentido, ir para a frente e para trás no mesmo trecho de estrada – é homotópica a uma das viagens regulares de ônibus. Além disso, viagens de ônibus com números diferentes não são homotópicas. A prova requer alguma técnica; a ideia básica é o número de rotação de Gauss, que conta o número total de vezes que a viagem dá a volta em torno do círculo no sentido anti-horário.[3] Ele nos mostra qual rota de ônibus da nossa viagem é homotópica.

Uma vez preenchidos os detalhes, esta descrição prova que o grupo fundamental de um círculo é o mesmo que o do grupo $\mathbb{Z}$ de inteiros para a adição. Para somar viagens basta somar seus números de rotação. A formiga podia usar esse invariante topológico para distinguir um universo circular de, por exemplo, uma reta infinita. Numa reta, qualquer viagem, por mais que fique ziguezagueando de um lado para outro, precisa em alguma etapa chegar a uma distância máxima de casa. Agora podemos encurtar a viagem continuamente encolhendo todas as distâncias de casa por um mesmo valor – primeiro para 99%, depois 98%, e assim por diante. Então, numa reta, *qualquer* viagem é homotópica a zero: ficar em casa. O

grupo fundamental da reta tem um único elemento: zero. Suas propriedades algébricas são triviais: 0 + 0 = 0. Portanto é chamado de grupo trivial, e como difere do grupo de todos os inteiros, a formiga pode saber a diferença entre viver numa reta e viver num círculo.

Como eu disse, há outros métodos, mas é assim que a formiga pode saber usando o grupo fundamental de Poincaré.

VAMOS CONTINUAR com o nosso trabalho de formiga (trocadilho intencional). Suponhamos agora que a formiga viva numa superfície. Mais uma vez, esse é o seu universo inteiro; ela não pode dar um passo para o lado e ver que tipo de superfície habita. Será que ela consegue descobrir a topologia do seu universo? Em particular, será que consegue saber a diferença entre uma esfera e um toro? Mais uma vez a resposta é "sim", e o método é o mesmo que usamos para um universo circular: entrar em um ônibus e fazer viagens circulares que comecem e terminem em casa. A viagem zero é "ficar em casa", o inverso de uma viagem é a mesma viagem no sentido oposto, e obtemos um grupo contanto que trabalhemos em classes de homotopia de viagens. Este é o grupo fundamental da superfície. Comparado com o universo circular há mais liberdade para criar viagens e deformá-las continuamente em outras viagens, mas a mesma ideia básica funciona.

O grupo fundamental é mais uma vez um invariante topológico, e a formiga pode usá-lo para descobrir se vive sobre uma esfera ou um toro. Se seu universo é uma esfera, então não importa que viagem a formiga faça, pode ser gradualmente deformada numa viagem zero: ficar em casa. Esse não é o caso se o universo é um toro. Algumas viagens podem ser deformadas até zero, mas uma viagem que passa uma vez pelo furo central, como na Figura 39 (esquerda), não pode. Essa afirmação necessita de prova, mas ela pode ser providenciada. No toro há viagens de ônibus padronizadas, mas agora os números dos ônibus são pares de inteiros (*m*, *n*). O primeiro número, *m*, especifica quantas vezes a viagem passa pelo furo central; o segundo número, *n*, especifica quantas vezes a viagem circunda o toro. A Figura 39 (direita) mostra a viagem (5, 2), que passa cinco vezes pelo furo central e circunda duas vezes o toro. Para somar viagens, somam-se os números correspondentes; por exemplo, (3, 6) + (2, 4) = (5, 10). O grupo fundamental do toro é o grupo $\mathbb{Z}^2$ de *pares* de inteiros.

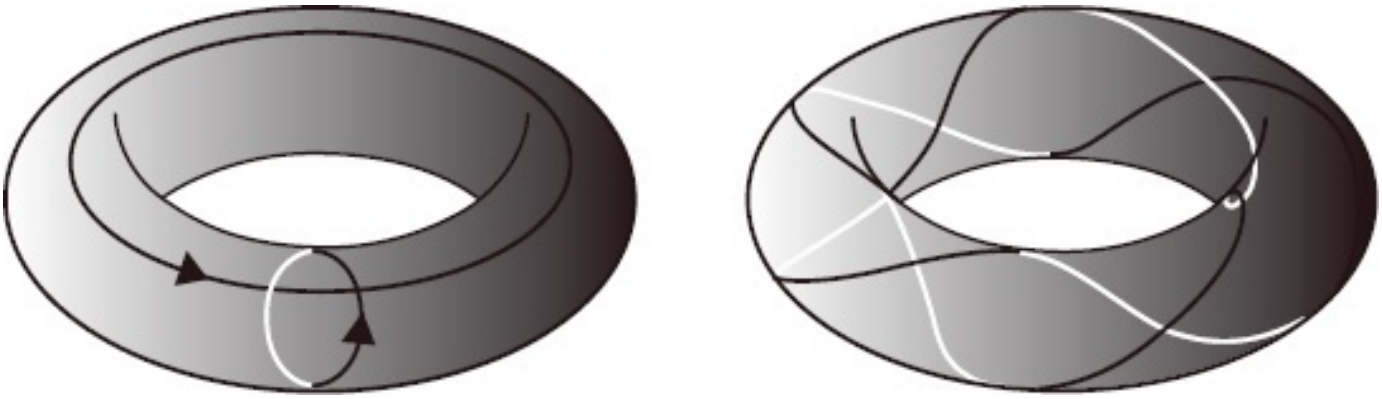


FIGURA 39 *Esquerda*: Viagens nos ônibus (1, 0) e (0, 1) sobre o toro. *Direita*: Viagem no ônibus (5, 2). Linhas claras estão no fundo.

Qualquer espaço topológico tem um grupo fundamental, definido exatamente da mesma maneira usando viagens – mais propriamente conhecidas como laços – que começam e terminam no mesmo ponto. Poincaré inventou o grupo fundamental para provar que seu espaço dodecaédrico não é uma 3-esfera, apesar de ter exatamente os mesmos invariantes homológicos. Sua receita original é belamente adaptada ao cálculo de seu grupo fundamental. A receita mais moderna de "recortar e colar" é ainda mais bem-adaptada. A resposta revela ser um grupo de 120 elementos relacionados com o dodecaedro. Em contraste, o grupo fundamental da 3-esfera tem apenas um elemento: o laço zero. Logo, o espaço dodecaédrico *não* é topologicamente equivalente a uma esfera, apesar de ter a mesma homologia, e Poincaré provou que sua afirmação de 1900 estava errada.

Poincaré seguiu adiante para especular sobre seu novo invariante: qual era o ingrediente que faltava numa caracterização topológica da 3-esfera? Talvez qualquer espaço 3-dimensional com o mesmo grupo fundamental que a 3-esfera – isto é, o grupo trivial – deva, na verdade, *ser* uma 3-esfera? Ele formulou esta sugestão de modo negativo como uma pergunta: "Consideremos uma variedade compacta 3-dimensional (espaço topológico) V sem fronteiras. É possível que o grupo fundamental de V possa ser trivial, mesmo que V não seja (topologicamente equivalente a) uma esfera 3-dimensional?" E deixou a questão em aberto, mas a própria crença plausível de que a resposta era óbvia – um "não" quando a questão é formulada dessa maneira – logo ficou conhecida como conjectura de Poincaré. E com a mesma rapidez tornou-se uma das mais notórias questões em aberto na topologia.

"GRUPO FUNDAMENTAL TRIVIAL" é outro modo de dizer "todo laço pode ser continuamente deformado até um ponto". Não é só uma 3-esfera que possui essa propriedade; o mesmo ocorre com qualquer *n*-esfera para qualquer dimensão *n*. Assim podemos fazer a mesma conjectura para uma esfera de qualquer dimensão. Essa afirmação é a conjectura de Poincaré *n*-dimensional. Ela vale para $n = 2$, pelo teorema de classificação para superfícies. E por cinquenta anos foi o mais longe que alguém conseguiu chegar.

Em 1961, Stephen Smale tomou emprestado um artifício da classificação de superfícies e o aplicou a dimensões superiores. Um modo de pensar num toro de *g* furos é começar com uma esfera e adicionar *g* alças – como a alça de uma xícara de chá ou de um jarro. Smale generalizou essa construção para qualquer número de dimensões, chamando o processo de decomposição em alças. Analisou como as alças podiam ser modificadas sem mudar a topologia do espaço e deduziu a conjectura de Poincaré em todas as dimensões maiores ou iguais a 7. Sua prova não funcionava para dimensões menores, mas outros matemáticos acharam um modo de repará-la: John Stallings para a dimensão 6 e Christopher Zeeman para a dimensão 5. Contudo, um passo vital, conhecido como artifício de Whitney, falhou para as dimensões 3 e 4 porque nestes espaços não há lugar suficiente para executar as manobras exigidas, e ninguém conseguiu achar substitutos efetivos. Veio à tona uma sensação generalizada de que a topologia, para espaços nessas duas dimensões, poderia ser incomum.

Essa sabedoria convencional foi abalada em 1982 quando Michael Freedman descobriu uma prova para a conjectura de Poincaré 4-dimensional que não requeria o artifício de Whitney. Era extremamente complicada, mas funcionava. Assim, depois de cinquenta anos com pouco progresso e vinte anos de atividade frenética, os topologistas haviam polido a conjectura de Poincaré em todas as dimensões, exceto aquela na qual Poincaré havia formulado originalmente a pergunta. Os sucessos foram impressionantes, mas os métodos usados para obtê-los forneciam muito pouco entendimento do caso 3-dimensional. Era necessário outro modo de pensar.

O que finalmente quebrou o impasse foi mais ou menos parecido com a tradicional lista de presentes de casamento: algo antigo, algo novo, algo emprestado… e, forçando um pouco a barra, algo azul. A antiga ideia era revisitar uma área da topologia que, depois de imensa atividade com espaços de dimensões superiores, acreditava-se ter se esgotado: a

topologia de superfícies. A nova ideia era repensar a classificação de superfícies de um ponto de vista que de início parecia completamente fora de propósito: a geometria clássica. A ideia emprestada era o fluxo de Ricci, que tomava sua motivação do formalismo matemático da teoria de Einstein da relatividade geral. A ideia azul não passava de uma especulação que se perdia no céu azul: algumas sugestões de longo alcance baseadas num rompante de intuição e um bocado de esperança.

Lembremos que superfícies orientáveis sem fronteira podem ser postas numa lista: cada uma é topologicamente equivalente a um toro com algum número de furos. Esse número é o genus da superfície, e quando é zero, a superfície é a esfera sem alças – isto é, a esfera. Esta simples palavra nos recorda que entre todas as esferas topológicas há uma superfície que sobressai como o arquétipo. Ou seja, a esfera unitária no espaço euclidiano. Esqueçamos, por um segundo, toda a história com folhas de borracha. Daqui a pouco a colocamos de volta. Concentre-se na velha e boa esfera euclidiana. Ela tem todo tipo de propriedades matemáticas adicionais, provenientes da geometria de Euclides. Suprema entre essas propriedades é a curvatura. Esta pode ser quantificada; em cada ponto de uma superfície geométrica há um número que mede o quanto a superfície é curva perto desse ponto. A esfera é a única superfície fechada no espaço euclidiano cuja curvatura é a mesma em qualquer ponto, e é positiva.

Isso é estranho, porque curvatura constante não é uma propriedade topológica. Mais esquisito ainda: a esfera não está sozinha. Existe também uma superfície geométrica padrão que se apresenta como o toro arquetípico. Ou seja, comecemos com um quadrado no plano, e consideremos coincidentes os lados opostos (Figura 12 no Capítulo 4). Quando desenhamos o resultado num espaço 3-dimensional, enrolando o quadrado para fazer seus lados coincidir, o resultado tem aparência curva. Mas do ponto de vista intrínseco, podemos trabalhar inteiramente com o quadrado e com as regras de colagem. Um quadrado tem uma estrutura geométrica natural: é uma região do plano euclidiano. O plano também tem curvatura constante, mas agora essa constante é *zero*. Um toro com essa geometria particular também tem curvatura zero, então é chamado de *toro plano*. O nome pode soar contraditório, mas para uma formiga habitando um toro plano, usando régua e transferidor para medir comprimentos e ângulos, a geometria local seria idêntica à do plano.

Os geômetras do século XVIII, tentando entender o axioma de Euclides acerca da existência de retas paralelas, dispostos a deduzir esse axioma

para o restante das premissas básicas de Euclides, fracassaram repetidamente, e acabaram percebendo que tal dedução não é possível. Há três tipos diferentes de geometria, cada uma obedecendo a toda condição requerida por Euclides, exceto o axioma das paralelas. Essas geometrias são chamadas euclidiana (o plano, onde o axioma das paralelas é válido), elíptica (a geometria sobre a superfície de uma esfera, com alguns enfeites, onde quaisquer duas linhas sempre se encontram e não existem paralelas) e hiperbólica (onde algumas linhas deixam de se encontrar, e as paralelas deixam de ser exclusivas). Além disso, a matemática clássica interpretava essas geometrias como geometrias de espaços curvos. A geometria euclidiana corresponde à curvatura zero, a geometria elíptica/esférica corresponde a uma curvatura constante positiva, e a geometria hiperbólica corresponde à curvatura constante negativa.

Acabamos de ver como obter as primeiras duas dessas geometrias: elas ocorrem na esfera e no toro plano. Em termos do teorema de classificação, são toros de g furos para $g = 0$ e 1. A única coisa que falta é a geometria hiperbólica. Será que todo toro de g furos tem uma estrutura geométrica natural, baseada em se tomar algum polígono no espaço hiperbólico e fazer coincidir alguns de seus lados? A resposta é surpreendente: "sim" para *qualquer* valor de g maior ou igual a 2. A Figura 40 mostra um exemplo para $g = 2$ baseado num octógono. Vou pular a geometria hiperbólica e a identificação dessa superfície como um 2-toro, mas isso pode ser demonstrado. Diferentes valores de g surgem quando tomamos polígonos diferentes, mas todo g aparece. Usando um jargão, um toro com dois ou mais furos tem uma estrutura hiperbólica natural. Logo, agora podemos reinterpretar a lista de superfícies-padrão:

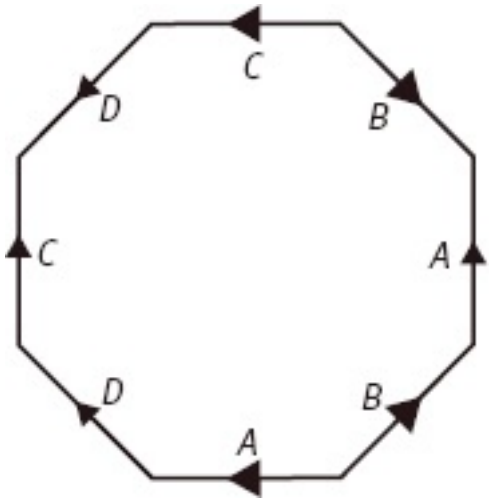


FIGURA 40 Fazendo um toro de dois furos a partir de um octógono, fazendo coincidir os lados com

os pares (*AA*, *BB*, *CC*, *DD*).

- Esfera, $g = 0$: geometria elíptica.
- Toro, $g = 1$: geometria euclidiana.
- Toro com g furos, $g = 2, 3, 4, \ldots$: geometria hiperbólica.

Pode parecer que jogamos fora o bebê junto com a água do banho, porque supostamente a topologia deve tratar da geometria em folha de borracha, não da geometria rígida. Mas agora podemos facilmente recolocar a borracha de volta. A geometria rígida é usada aqui apenas para *definir* as superfícies-padrão. Ela fornece descrições simples, que ocorrem ter estrutura rígida extra. Agora, vamos amolecer essa rigidez – vamos, efetivamente, permitir que o espaço *se torne* uma borracha. Permitir que ele se deforme de maneiras proibidas pela rigidez. Agora obtemos superfícies que são topologicamente equivalentes às superfícies-padrão, mas não são equivalentes em termos de movimentos rígidos. O teorema da classificação nos diz que toda superfície topológica pode ser obtida dessa maneira.

OS TOPOLOGISTAS TINHAM consciência desse elo entre a geometria e o teorema de classificação para superfícies, mas dava impressão de uma coincidência engraçada, sem dúvida uma consequência das possibilidades bastante limitadas em duas dimensões. Todo mundo sabia que o caso 3-dimensional era muito mais rico e, em particular, espaços de curvatura constante não esgotavam as possibilidades. Foi preciso um dos maiores geômetras do mundo, William Thurston, para perceber que a geometria rígida ainda poderia ser relevante para a topologia 3-dimensional. Já havia alguns indícios: a 3-esfera de Poincaré tem uma geometria elíptica/esférica natural proveniente de sua definição. Embora um dodecaedro padrão habite um espaço euclidiano, o ângulo entre faces adjacentes é menor que 120 graus, então três desses ângulos não completam um círculo inteiro. Para remediar isso, temos que inflar o dodecaedro de modo que suas faces fiquem ligeiramente arqueadas: isso torna a geometria natural esférica, não euclidiana. Analogamente, triângulos sobre uma esfera também ficam arqueados. O 3-toro, obtido ao se coincidir faces opostas de um cubo, tem uma geometria plana – isto é, euclidiana, exatamente como seu análogo 2-dimensional. Max Dehn e outros descobriram alguns espaços topológicos 3-

dimensionais com geometrias hiperbólicas naturais.

Thurston começou a ver indícios de uma teoria geral, mas foram necessárias duas inovações para torná-la remotamente plausível. Primeira, a gama das geometrias 3-dimensionais precisava ser estendida. Thurston anotou condições razoáveis, e provou que exatamente oito geometrias as satisfaziam. Três delas são as clássicas: esférica, euclidiana e hiperbólica. Outras duas são como cilindros: planas numa direção, curvas nas outras duas direções. A parte curva pode ser curvada positivamente, a 2-esfera, ou negativamente, o plano hiperbólico. Finalmente, há outras três geometrias, bastante técnicas.

Segunda: alguns espaços 3-dimensionais não sustentavam qualquer uma das oito geometrias. A resposta era cortar o espaço em pedaços. Um deles podia ter estrutura geométrica esférica, outro hiperbólica, e assim por diante. Para ser útil, o corte tinha de ser feito de maneira muito precisamente constrita, de modo que a remontagem dos pedaços fornecesse informação proveitosa. A boa notícia era que, em muitos exemplos, isso acabou se revelando possível. Em 1982, num grande salto de imaginação, Thurston enunciou sua conjectura de geometrização: *todo* espaço 3-dimensional pode ser cortado, de maneira essencialmente exclusiva, em pedaços, cada um com uma estrutura geométrica correspondente a uma das oito geometrias possíveis. Também provou que se sua conjectura de geometrização fosse verdadeira, então a conjectura de Poincaré seria uma simples consequência.

ENTRETANTO, uma segunda linha de ataque estava surgindo, também geométrica, também baseada em curvatura, mas vindo de uma área muito diferente: a física matemática. Gauss, Riemann e uma escola de geômetras italianos haviam desenvolvido uma teoria geral de espaços curvos, chamada variedades, com um conceito de distância que estendia enormemente a geometria clássica euclidiana e não euclidiana. A curvatura não precisava mais ser constante: podia variar suavemente de um ponto a outro. Uma forma como a de um osso de cachorro, por exemplo, é curvada positivamente em cada uma das extremidades, mas negativamente no meio, e o grau de curvatura varia suavemente de uma região para outra. A curvatura é quantificada usando dispositivos matemáticos conhecidos como tensores. Por volta de 1915, Albert Einstein percebeu que tensores de curvatura eram exatamente o que ele necessitava para estender sua teoria

da relatividade especial, que tratava de espaço e tempo, para a relatividade geral, que também incluía a gravidade. Em sua teoria, o campo gravitacional é representado como a curvatura do espaço, e as equações de campo de Einstein descrevem como a medida associada à curvatura, o tensor de curvatura, varia em resposta à distribuição da matéria. Com efeito, a curvatura do espaço *flui* com o passar do tempo; o universo, ou alguma parte dele, muda espontaneamente de forma.

Richard Hamilton, um especialista em geometria riemanniana, deu-se conta de que o mesmo recurso podia ser aplicado de modo mais geral, podendo levar a uma prova da conjectura de Poincaré. A ideia era trabalhar com uma das medidas mais simples de curvatura, chamada curvatura de Ricci, em homenagem ao geômetra italiano Gregorio Ricci-Curbastro. Hamilton escreveu uma equação que especificava como a curvatura de Ricci deveria variar com o tempo: o fluxo de Ricci. A equação era montada de modo tal que a curvatura deveria se redistribuir gradualmente da forma mais regular possível. Isto é, um pouco como o gato sob o tapete no Capítulo 4, mas agora, mesmo que o gato não possa fugir, ele pode se espalhar numa camada regular. (Aqui é essencial um gato topológico.)

Por exemplo, no caso 2-dimensional, comecemos com uma superfície em forma de pera (Figura 41). Ela possui uma região numa das extremidades que tem acentuada curvatura positiva; uma região na outra extremidade, mais gorda, também com curvatura positiva, mas não tão acentuada; e uma faixa intermediária onde a curvatura é negativa. O fluxo de Ricci efetivamente transporta a curvatura da extremidade acentuadamente curvada (e, em menor proporção, também da outra extremidade) para a faixa negativamente curvada, até toda a curvatura negativa ter sido compensada. Nessa fase o resultado é uma superfície toda convexa, com curvatura positiva em todo lugar. O fluxo de Ricci continua a redistribuir a curvatura, tirando-a de regiões altamente curvadas e passando-a para regiões menos curvadas. Com o tempo aumentando, a superfície vai se aproximando mais e mais daquela que tem curvatura positiva constante – ou seja, a esfera euclidiana. A topologia permanece a mesma, ainda que a forma detalhada mude, e assim, acompanhando o fluxo de Ricci, podemos provar que a superfície inicial em forma de pera é topologicamente equivalente a uma esfera.

Nesse exemplo, o tipo topológico da superfície era óbvio para se começar, mas a mesma estratégia geral funciona para qualquer variedade. Inicie com uma forma complicada e acompanhe o fluxo de Ricci. Com o

passar do tempo, a curvatura se redistribui mais regularmente, e a forma torna-se mais simples. Em última instância, você deve terminar com a forma mais simples com a mesma topologia que a variedade original, qualquer que seja ela. Em 1981, Hamilton provou que essa estratégia funciona em duas dimensões, fornecendo uma nova prova para o teorema de classificação para superfícies.

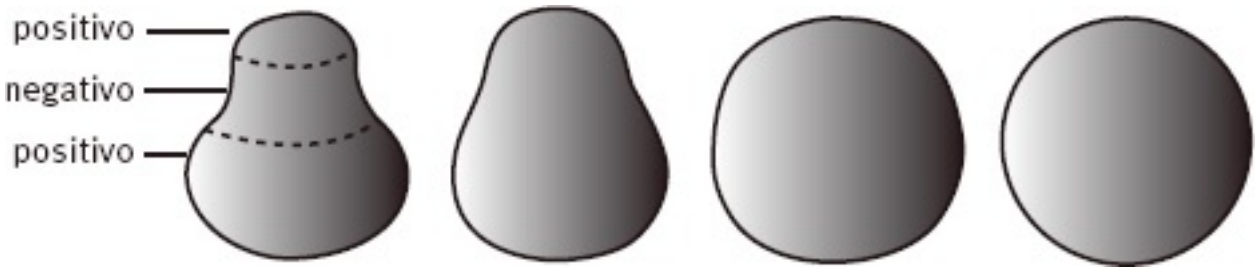


FIGURA 41 Como o fluxo de Ricci transforma uma pera em uma esfera.

Ele também fez significativo progresso na estratégia análoga para variedades 3-dimensionais, mas aí havia um sério obstáculo. Em duas dimensões, toda superfície automaticamente se simplifica seguindo o fluxo de Ricci. O mesmo ocorre em três dimensões se a variedade inicial tiver curvatura estritamente positiva em cada ponto: nunca zero ou negativa. Infelizmente, se houver pontos onde a curvatura é zero, e frequentemente há, o espaço pode se emaranhar em si mesmo à medida que flui. Isso cria singularidades: locais onde a variedade deixa de ser suave. Nesses pontos, a equação para o fluxo de Ricci cai por terra, e a redistribuição da curvatura precisa parar. A maneira natural de contornar esse obstáculo é compreender o aspecto dessas singularidades e redesenhar a variedade – talvez cortando-a em pedaços – para que o fluxo de Ricci possa receber um impulso inicial. Contanto que se tenha suficiente controle de como a topologia da variedade remodelada relaciona-se com a da variedade original, esta estratégia modificada pode ser bem-sucedida. Infelizmente, Hamilton também percebeu que para espaços 3-dimensionais, as singularidades no fluxo de Ricci podem ser realmente muito complicadas – aparentemente complicadas demais para usar esse tipo de subterfúgio. O fluxo de Ricci logo se tornou uma técnica-padrão em geometria, mas fracassou em provar a conjectura de Poincaré.

No ano 2000, os matemáticos ainda não haviam solucionado essa conjectura, e sua importância passou a ser reconhecida quando foi considerada um dos sete problemas do milênio. A essa altura também tinha

ficado claro que se a ideia de Hamilton pudesse ser posta para funcionar com suficiente generalidade, isso simplesmente resultaria na conjectura de Poincaré. E provaria também a conjectura de geometrização de Thurston. O prêmio era reluzente, mas permanecia assustadoramente longe do alcance.

A MATEMÁTICA É como os outros ramos da ciência: para uma pesquisa ser aceita como correta, precisa ser publicada, e para isso acontecer necessita sobreviver à revisão dos pares. Especialistas no campo em questão precisam ler com cuidado o artigo, conferir a lógica e assegurar que os cálculos estejam corretos. Esse processo pode levar um tempo longo para uma peça matemática importante e complicada. Como mencionei no Capítulo 1, o remédio costumava ser um pré-impresso, mas nos dias de hoje há um website padrão, o arXiv, onde podem ser postados pré-impressos eletrônicos, sujeitos a um processo de arbitragem parcial e a procedimentos de endosso para retirada de "entulho". Hoje, os pesquisadores encontram novos resultados primeiro no arXiv ou no próprio site do autor.

Em 2002, Grigori Perelman postou um pré-impresso no arXiv sobre o fluxo de Ricci. Ele fazia uma alegação extraordinária: o fluxo é como um gradiente. Quer dizer, há um sentido "descendente" bem-definido, uma única grandeza numérica associada com a forma da variedade, e a variedade flui para baixo no sentido de que esta grandeza sempre diminui com o passar do tempo. É análogo a um pico numa paisagem e provê uma medida quantitativa do que significa "simplificar" uma variedade. Fluxos do tipo gradiente são bastante restritos: não podem ficar dando voltas e voltas nem se comportar caoticamente. Ninguém parece ter suspeitado de que o fluxo de Ricci seria tão manso. Mas Perelman não se limitou a fazer a alegação: ele a provou. E terminou esboçando um argumento que provaria a conjectura de geometrização de Thurston – que implica a conjectura de Poincaré, mas vai muito além – prometendo maiores detalhes em postagens subsequentes no arXiv. Nos oito meses seguintes ele postou dois artigos correlatos contendo muitos dos detalhes prometidos.

A primeira postagem causou bastante rebuliço. Perelman alegava ter executado todo o programa de Hamilton, usando o fluxo de Ricci para simplificar uma variedade 3-dimensional e provar que o resultado era exatamente o que Thurston predissera. As outras duas postagens acrescentaram peso adicional à sensação de que Perelman sabia do que

estava falando, e que suas ideias iam bem além de esboçar uma estratégia plausível com estranha lacuna lógica ou premissa não provada. O ceticismo habitual da comunidade matemática com respeito a alegações de resolução de um grande problema ficou emudecido; havia uma sensação geral de que ele podia muito bem ter conseguido.

No entanto, o diabo mora no detalhe, e em matemática o detalhe pode ser realmente diabólico. O trabalho precisava ser verificado, em extensão e profundidade, por gente que compreendesse as áreas envolvidas e estivesse ciente das armadilhas potenciais. E não era algo simples e direto, porque Perelman combinara pelo menos quatro áreas muito diferentes de matemática e física matemática, e pouca gente entendia mais do que uma ou duas delas. Decidir se a prova estava correta ou não exigiria um bocado de trabalho em equipe e muito esforço. Além disso, os pré-impressos no arXiv não incluíam todos os detalhes em nível normal para um artigo publicado. Para pré-impressos estavam redigidos de forma bastante clara, mas nem sempre tinham os pingos nos is. Assim, os entendedores precisavam reconstituir boa parte do raciocínio de Perelman – e ele estivera profundamente imerso no trabalho durante anos.

Tudo isso levou tempo. Perelman dava aulas sobre sua prova e respondia a e-mails questionando vários passos. Sempre que alguém encontrava algo que parecia ser uma lacuna, ele respondia rapidamente com explicações adicionais, preenchendo-a. Os sinais eram encorajadores. Mas ninguém estava disposto a arriscar sua reputação declarando publicamente que Perelman provara a conjectura de Poincaré, muito menos a conjectura mais difícil da geometrização, até ter confiança de que não havia erros na prova. Assim, apesar da opinião genericamente favorável ao trabalho de Perelman, a aceitação pública foi a princípio retida. Era algo inevitável, mas também desafortunado, pois à medida que a espera se arrastava, Perelman foi ficando cada vez mais irritado pelo que parecia ser uma postura "em cima do muro". Ele *sabia* que sua prova estava correta. Ele a compreendia tão bem que não conseguia ver por que os outros estavam tendo dificuldades. Declinou redigir o trabalho em mais detalhes ou submetê-lo a uma publicação científica. No que lhe dizia respeito, era assunto acabado, e os pré-impressos do arXiv continham tudo que era exigido. Perelman parou de responder a perguntas a respeito de detalhes que alegadamente faltavam. Para ele, não faltava nada. Vamos lá, rapazes, vocês podem resolver isso sem mais ajuda minha. Não é tão difícil assim.

Alguns relatos têm sugerido que sob esse aspecto a comunidade matemática foi injusta com Perelman. Mas essa é uma visão que entende mal como a comunidade matemática funciona quando grandes problemas são alegadamente solucionados. Teria sido irresponsável simplesmente dar-lhe tapinhas nas costas, dizendo "muito bem", e ignorar os passos que faltavam em seus pré-impressos. Foi inteiramente apropriado, na verdade inevitável, pedir-lhe que preparasse tratamentos mais extensivos, adequados para publicação. Num problema de tal importância, um trabalho apressado é perigoso e inaceitável. Os conhecedores da matéria deixaram suas coisas de lado para passar um bom tempo na prova de Perelman, e mantiveram seu ceticismo natural ao largo numa proporção incomum. O tratamento que lhe deram, se houve alguma diferença, foi *mais* favorável do que o normal. E no final, quando o processo foi completado, o trabalho foi aceito como correto.

A essa altura, porém, Perelman tinha perdido a paciência. Talvez não tenha ajudado muito o fato de ter resolvido um problema de tal importância que nada mais poderia lhe ser equiparável. Ele era como um alpinista que escalara o Everest sozinho, sem oxigênio. Não restavam desafios comparáveis. A publicidade na mídia o repugnava: ele queria a aceitação de seus pares, não de apresentadores de televisão. Assim, não foi nenhuma grande surpresa que, quando seus pares finalmente concordaram que ele estava certo e lhe ofereceram a medalha Fields e o prêmio Clay, Perelman não quis nem saber.

A prova de Perelman é profunda e elegante, e abre um mundo novo para a topologia. Ela implementa o programa do fluxo de Ricci elaborado por Hamilton encontrando maneiras inteligentes de contornar a ocorrência de singularidades. Uma delas é mudar as escalas de espaço e tempo para se livrar da singularidade. Quando esta abordagem falha, diz-se que a singularidade colapsa. Em tais casos, ele analisa a geometria do fluxo de Ricci em detalhe, classificando a maneira como o colapso pode ocorrer. Efetivamente, o espaço estende tentáculos cada vez mais finos, talvez em profusão, como galhos de uma árvore. Sempre que um tentáculo está perto o suficiente de colapsar, ele pode ser cortado, e o toco pontudo, agudamente curvo, pode ser tirado e substituído por uma cápsula suave. Para alguns desses tentáculos, o fluxo de Ricci se tritura até parar: se isso acontecer, deixe de lado. Caso contrário, o fluxo de Ricci pode ser reiniciado. Logo, alguns tentáculos terminam em cápsulas suaves, e outros são temporariamente interrompidos, mas continuam a fluir.

Esse procedimento de “recortar e colar” fatias cobre o espaço mais ou menos da mesma maneira que a dissecção em pedaços de Thurston, cada um com suas oito geometrias, e os dois procedimentos acabam dando mais ou menos os mesmos resultados. Um ponto técnico é vital: as operações de poda não se acumulam cada vez mais rápido, de modo que infinitas delas ocorrem num tempo finito. Essa é uma das partes mais complicadas da prova.

ALGUNS COMENTARISTAS TÊM criticado a comunidade matemática por tratar Perelman injustamente. Ninguém deve ser imune a críticas, e houve alguns incidentes que podem muito bem ser classificados como injustos ou falta de consideração, mas a comunidade matemática reagiu de forma rápida e positiva ao trabalho de Perelman. E também reagiu com cautela, que é absolutamente o padrão em matemática e ciência, por excelentes motivos. O brilho inevitável da publicidade, amplificado pelo prêmio de 1 milhão de dólares, causou impacto em todo mundo, inclusive em Perelman.

Desde a primeira postagem de Perelman no arXiv em novembro de 2002 até o anúncio, em março de 2010, de que lhe fora concedido o prêmio Clay, passaram-se oito anos. No entanto, aquela primeira postagem atacou apenas parte do problema. A maior parte do restante foi postada em março de 2003. Em setembro de 2004, dezoito meses depois da segunda postagem, as comunidades topológica e do fluxo de Ricci já haviam examinado a prova – processo este que teve início apenas alguns dias após o primeiro post – e os principais peritos anunciaram que “entenderam a prova”. Tinham achado erros, tinham achado lacunas, mas estavam convictos de que as falhas podiam ser sanadas. Dezoito meses são, na verdade, um prazo notavelmente rápido quando algo tão importante está em jogo.

No fim de 2005, a União Matemática Internacional abordou Perelman e lhe ofereceu a medalha Fields, a mais alta honraria no campo, a ser concedida no Congresso Internacional de Matemáticos em 2006. Esse congresso é organizado a cada quatro anos, então era a primeira oportunidade de reconhecer seu trabalho. Como restavam algumas dúvidas a respeito da prova completa da conjectura de Poincaré – ainda apareciam erros – a medalha foi oficialmente concedida por progressos na compreensão do fluxo de Ricci, a parte dos pré-impressos de Perelman que agora eram consideradas livres de erros.

As condições para a concessão do prêmio estão apresentadas no site do Instituto Clay. Em particular, uma proposta de solução precisa ser publicada numa revista científica reconhecida e ainda ser aceita pela comunidade matemática dois anos depois. Em seguida, um comitê consultor especial examina o assunto e recomenda se o prêmio deve ser concedido ou não. Perelman não atendera à primeira condição, e não parece provável que algum dia venha a atender. Em sua opinião, os pré-impressos no arXiv bastam. Não obstante, o Instituto Clay deixou de lado essa exigência e deu início à espera estatutária de dois anos para ver se surgiam quaisquer outros erros ou questões. Isso terminou em 2008, após o que os procedimentos do instituto, cuidadosamente estruturados para evitar conceder o prêmio prematuramente, precisaram ser seguidos.

É verdade que alguns peritos foram lentos em manifestar sua crença de que a prova estava correta. A razão é simples: estavam genuinamente incertos. Não é um grande exagero dizer que a única pessoa capaz de captar rapidamente a prova de Perelman era outro Perelman. Não se pode ler uma prova matemática como um músico lê uma partitura. Você precisa se convencer de que tudo faz sentido. Sempre que um argumento fica muito complicado, você sabe que ali há uma grande chance de erro. O mesmo se aplica quando as ideias ficam simples demais; mais do que uma prova em potencial foi vítima de alguma afirmação tão evidente que nenhuma prova parecia necessária. Até os peritos estarem genuinamente seguros de que a prova estava basicamente correta – ponto este em que deram pleno crédito a Perelman *apesar* dos erros e lacunas remanescentes – era sensato suspender o julgamento. Basta pensar em toda a balbúrdia em torno do trabalho sobre fusão a frio que acabou sendo desacreditado. Cautela é a resposta profissional correta, e o chavão se aplica: alegações extraordinárias requerem evidência extraordinária.

Por que Perelman rejeitou a medalha Fields e declinou do prêmio Clay? Só ele sabe, mas não estava interessado nesse tipo de reconhecimento e o disse repetidas vezes. Ele já recusara prêmios menores. Deixou claro desde o início que não queria publicidade prematura; ironicamente, é a mesma razão pela qual os peritos estavam compreensivelmente relutantes em dar o mergulho cedo demais. Para ser realista, não havia a menor chance de a mídia *não* notar seu trabalho. Durante anos, a comunidade matemática vem fazendo um grande esforço para os jornais, o rádio e a televisão se interessarem no assunto. Não faz muito sentido queixar-se quando o esforço dá resultado, ou esperar que a mídia ignore a história matemática

mais quente desde o último teorema de Fermat. Mas Perelman não via as coisas dessa maneira, e recolheu-se na sua concha. Há uma oferta na mesa para disponibilizar o prêmio para propósitos educacionais e outros, se ele concordar. Até agora ele não deu resposta.

11. Não podem ser todos fáceis

O problema P/NP

ATUALMENTE, é comum os matemáticos usarem computadores para resolver problemas, mesmo problemas grandes. Computadores são bons em aritmética, mas a matemática vai muito além de meras "somas", então introduzir um problema no computador raramente é algo simples. Muitas vezes a parte mais difícil do trabalho é converter o problema em outro que um cálculo de computador possa resolver, e mesmo assim o computador pode ter que enfrentar batalhas. Muitos dos grandes problemas que foram resolvidos recentemente envolvem pouco ou nenhum trabalho de computador. O último teorema de Fermat e a conjectura de Poincaré são exemplos.

Quando computadores foram usados para resolver grandes problemas, como o problema das quatro cores ou a conjectura de Kepler, a máquina efetivamente desempenha o papel de um criado. Mas às vezes os papéis se invertem, com a matemática servindo de criada da ciência da computação. A maior parte do trabalho dos primeiros tempos em projetos de computador fez bom uso de sacações matemáticas, por exemplo, a ligação entre a álgebra booleana – uma formulação algébrica da lógica – e troca de circuitos, desenvolvida em particular pelo engenheiro Claude Shannon, inventor da teoria da informação. Hoje, tanto os aspectos práticos como teóricos do computador dependem do uso extensivo da matemática, em muitas áreas diferentes.

Um dos problemas do milênio do Instituto Clay reside na região limítrofe entre a matemática e a ciência da computação. E pode ser vista dos dois modos: a ciência da computação como serva da matemática e a matemática como serva da ciência da computação. O que ela requer, e está ajudando a criar, é algo mais equilibrado: uma parceria. O problema trata de algoritmos de computador, os esqueletos matemáticos a partir dos quais são feitos os programas. O conceito crucial aqui é o grau de eficiência do algoritmo: quantos passos computacionais são necessários para obter uma resposta para uma determinada quantidade de dados alimentados. Em termos práticos, isso nos diz quanto tempo o computador vai demorar para

resolver um problema de determinado tamanho.

A palavra algoritmo remonta à Idade Média, quando Muhammad ibn Mūsā al-Khwārizmī escreveu um dos primeiros livros sobre álgebra. Antes dele, Diofanto introduzira um elemento que associamos com álgebra: símbolos. No entanto, usava os símbolos como abreviaturas, e seus métodos de resolver equações eram apresentados por meio de exemplos específicos – embora comuns. Onde atualmente escreveríamos algo como "$x + a = y$, portanto $x = y - a$", Diofanto escreveria "suponhamos $x + 3 = 10$, então $x = 10 - 3 = 7$", esperando que os leitores entendessem que a mesma ideia funcionaria se 3 e 10 fossem substituídos por outros números. Ele explicava seu exemplo ilustrativo usando símbolos, mas não manipularia os símbolos como tais. Al-Khwārizmī tornou explícita a receita genérica. E o fez usando palavras, não símbolos, mas tinha a ideia básica, e costuma ser considerado o pai da álgebra. Na verdade, esse nome vem do título de seu livro: *al-Kitāb al-Mukhtasar fī Hisāb al-Jabr wa'l-Muqābala* (O livro compêndio sobre cálculo por complementação e balanceamento). *Al-jabr* virou álgebra. A palavra "algoritmo" vem de uma versão medieval de seu nome, *Algorismus*, e é agora usado para significar um processo matemático específico para resolver problemas, um processo que garanta encontrar uma solução com a condição de que você espere o tempo necessário.

Tradicionalmente, os matemáticos consideravam um problema solucionado se, a princípio, pudessem escrever um algoritmo que levasse a uma resposta. Raramente usavam essa palavra, preferindo apresentar, digamos, uma fórmula para a solução, que é um tipo específico de algoritmo em linguagem simbólica. Não era de importância crucial que fosse possível aplicar a fórmula na prática: ela *era* a solução. Mas o uso de computadores mudou essa visão, porque fórmulas que haviam sido complicadas demais para cálculos à mão podiam se tornar práticas com a ajuda de um computador. Contudo, era um pouco decepcionante descobrir, como às vezes ocorria, que a fórmula continuava sendo muito complicada: embora o computador pudesse tentar rodar o algoritmo, era lento demais para chegar a uma resposta. Assim, a atenção voltou-se para encontrar algoritmos eficientes. Tanto matemáticos como cientistas da computação tinham grande interesse em desenvolver algoritmos que realmente dessem respostas em um período de tempo razoável.

Dado um algoritmo, é relativamente simples estimar quanto tempo vai levar (medido pelo número de passos computacionais exigidos) para resolver um problema com um determinado volume de dados de entrada,

ou *input*. Isso pode requerer uma certa técnica, mas você sabe qual é o processo envolvido e sabe muita coisa sobre o que está sendo feito. É muito mais difícil projetar um algoritmo eficiente se aquele com o qual você começa se revela ineficiente. E é mais difícil ainda decidir quanto o algoritmo mais eficiente para um dado problema pode ser bom ou ruim, porque isso envolve contemplar todos os algoritmos possíveis, e você não sabe quais são eles.

O trabalho inicial sobre essas questões levou a uma dicotomia rude, porém conveniente, entre algoritmos que eram eficientes, num sentido imediato e grosseiro, e os que não o eram. Se a duração do processo cresce com relativa lentidão à medida que aumenta o tamanho do *input*, o algoritmo é eficiente e o problema é fácil. Se a duração do processo cresce cada vez mais rápido à medida que aumenta o tamanho do *input*, o algoritmo é ineficiente e o problema é difícil. A experiência nos diz que, embora alguns problemas sejam fáceis, nesse sentido, a maioria parece ser difícil. De fato, se todos os problemas matemáticos fossem fáceis, os matemáticos não teriam emprego. O problema do prêmio do milênio pede uma prova rigorosa de que existe pelo menos um problema difícil – ou que, contrariando a experiência, todos os problemas são fáceis. É conhecido como problema P/NP, e ninguém tem a menor pista de como resolvê-lo.

JÁ ENCONTRAMOS uma medida de eficiência aproximada no Capítulo 2. Um algoritmo é classe P se tem tempo de processamento polinomial. Em outras palavras, o número de passos necessários para obter uma resposta é proporcional a alguma potência fixa, tal como um quadrado ou um cubo, do volume de dados no *input*. Tais algoritmos são eficientes, falando de forma bastante ampla. Se o *input* é um número, o tamanho é a sua quantidade de dígitos, não o número em si. A razão é que a quantidade de informação necessária para especificar o número é o espaço que ele ocupa na memória do computador, que é (proporcional à) quantidade de dígitos. Um problema é classe P se existe um algoritmo classe P que o solucione.

Qualquer outro algoritmo ou problema pertence à classe não-P, e a maioria é ineficiente. Entre eles estão aqueles cujo tempo de processamento é exponencial aos dados de entrada: aproximadamente igual a um número fixo elevado à potência do tamanho do *input*. Estes são classe E, e são definitivamente ineficientes.

Alguns algoritmos são tão eficientes que rodam muito mais depressa

que o tempo polinomial. Por exemplo, para determinar se um número é par ou ímpar, olhe o seu último dígito. Se (em notação decimal) for 0, 2, 4, 6 ou 8, o número é par; se não, é ímpar. O algoritmo tem no máximo seis passos:

O último dígito é 0? Caso seja, então PARE. O número é par.

O último dígito é 2? Caso seja, então PARE. O número é par.

O último dígito é 4? Caso seja, então PARE. O número é par.

O último dígito é 6? Caso seja, então PARE. O número é par.

O último dígito é 8? Caso seja, então PARE. O número é par.

PARE. O número é ímpar.

Logo, o tempo de processamento é no máximo 6, não importa o tamanho do *input*. Ele pertence à classe "tempo constante".

Colocar uma lista de palavras em ordem alfabética é um problema classe P. Um modo simples de realizar tal tarefa é o tipo bolha, que recebe este nome porque efetivamente as palavras vão subindo na lista como bolhas num copo de refrigerante gasoso caso estejam numa posição mais baixa na lista do que a ordem alfabética requer. O algoritmo trabalha repetidamente ao longo da lista, comparando palavras adjacentes, e as troca de lugar se estiverem na ordem errada. Por exemplo, vamos supor que a lista comece com

PORCO MACACO GATO CACHORRO

Ao correr a lista pela primeira vez temos

MACACO PORCO GATO CACHORRO
MACACO **GATO PORCO** CACHORRO
MACACO GATO **CACHORRO PORCO**

onde as palavras em negrito são as que acabaram de ser comparadas. Na segunda passada temos:

GATO MACACO CACHORRO PORCO

GATO **CACHORRO MACACO** PORCO
GATO CACHORRO **MACACO PORCO**

A terceira passada fica:

CACHORRO GATO MACACO PORCO
CACHORRO **GATO MACACO** PORCO
CACHORRO GATO **MACACO PORCO**

Na quarta passada, nada se move, então sabemos que terminamos. Note como cachorro vai subindo passo a passo como uma bolha até o alto (ou seja, a frente).

Com quatro palavras, o algoritmo corre por três comparações em cada etapa, e há quatro etapas. Com n palavras, há $n - 1$ comparações por etapa e n etapas, um total de $n(n - 1)$ passos. Isso é um pouco menos que n^2, então o tempo de processamento é polinomial, na verdade, quadrático. O algoritmo pode terminar antes, mas no pior dos casos, quando as palavras estiverem exatamente na ordem inversa, são necessários $n(n - 1)$ passos. A seleção tipo bolha é óbvia e é classe P, mas está longe de ser o algoritmo de seleção mais eficiente. O tipo mais rápido de comparação, que é montado de maneira mais inteligente, percorre $n \ln n$ passos.

Um algoritmo simples com tempo de processamento exponencial, classe E, é "imprima uma lista de todos os números binários com n dígitos". Há 2^n números na lista, e imprimir cada um (e calculá-lo) leva aproximadamente n passos, de modo que o tempo de processamento é aproximadamente $2^n n$, que é maior que 2^n porém menor que 3^n quando n é suficientemente grande. Todavia, esse exemplo é um pouquinho bobo porque o que o deixa tão lento é o tamanho da resposta, o *output*, não a complexidade do cálculo, e essa observação vai se revelar crucial mais adiante.

Um algoritmo classe E mais comum resolve o problema do caixeiro-viajante. Um vendedor precisa visitar um número de cidades. Ele pode fazê-lo em qualquer ordem. Qual é a rota que visita todas elas na menor distância total? O modo ingênuo de resolver essa questão é: listar todas as rotas possíveis, calcular a distância total para cada uma e encontrar a menor delas. Com n cidades há

$$n! = n \times (n - 1) \times (n - 2) \times \ldots 3 \times 2 \times 1$$

rotas (leia-se "*n* fatorial"). Isso cresce mais depressa que qualquer exponencial.[1] Um método mais eficiente, chamado programação dinâmica, resolve o problema do caixeiro-viajante em tempo exponencial. O primeiro método desses, o algoritmo Held-Karp, encontra o trajeto mais curto em 2^n n^2 passos, o que novamente está entre 2^n e 3^n quando *n* é grande o bastante.

Mesmo que esses algoritmos sejam "ineficientes", podem ser usados artifícios especiais para encurtar a computação quando o número de cidades é grande para os padrões humanos, mas não grande demais para que os artifícios deixem de ser efetivos. Em 2006, D.L. Applegate, R.M. Bixby, V. Chvátal e W.J. Cook solucionaram o problema do caixeiro-viajante para 85.900 cidades, e este ainda era o recorde em meados de 2012.[2]

ESSES EXEMPLOS DE ALGORITMOS não se limitam a ilustrar o conceito de eficiência. Eles ressaltam o meu ponto de vista em relação à dificuldade de encontrar algoritmos melhores, e a dificuldade ainda maior de achar um que seja o mais eficiente possível. Todos os algoritmos conhecidos para o problema do caixeiro-viajante são classe E, tempo exponencial – mas isso não significa que não exista algoritmo eficiente. Simplesmente mostra que ainda não achamos um. Há duas possibilidades: não encontramos um algoritmo melhor porque não somos espertos o bastante, ou não achamos um algoritmo melhor porque não existe.

O Capítulo 2 é um caso a analisar. Até a equipe de Agrawal encontrar seu algoritmo classe P para testar números primos, o melhor algoritmo conhecido era não-P. E mesmo assim era bastante bom, com tempo de processamento *n* ln *n* para números de *n* dígitos, o que na verdade é melhor que o algoritmo de Agrawal-Kayal-Saxena até chegarmos a números com $10^{1.000}$ dígitos. Antes de seu algoritmo ser descoberto, as opiniões sobre a situação do teste de primalidade se dividiam. Alguns conhecedores desconfiavam que era classe P e que um algoritmo conveniente seria encontrado; outros pensavam que não era. O novo algoritmo veio do nada, uma entre mil ideias que alguém poderia ter tentado; aconteceu de essa dar certo. O precedente aqui nos obriga a ter os pés no chão: não sabemos, não podemos adivinhar, e os melhores palpites dos conhecedores podem ser bons ou não.

O grande problema que nos diz respeito aqui pede a resposta a uma pergunta mais fundamental: existe algum problema difícil? Será que todos podem ser fáceis, só que não somos espertos o bastante? O enunciado real

é mais sutil, porque já vimos um exemplo de problema que é indubitavelmente difícil: imprimir uma lista de todos os números binários com *n* dígitos. Conforme ressaltei, isso é um pouco tolo: a dificuldade não reside nos cálculos, mas no simples mecanismo de imprimir uma resposta muito longa. Sabemos que não há atalho, porque a resposta é tão comprida *por definição*. Se fosse mais curta, não seria a resposta.

Para apresentar uma questão coerente, exemplos triviais como esse precisam ser eliminados. A maneira de fazê-lo é introduzir outra classe de algoritmo, a classe NP. Esta não é a classe não-P; é a classe de algoritmos processados em tempo polinomial não determinístico. O jargão significa que, por mais que o algoritmo demore a vir com sua resposta, podemos *verificar que a resposta está certa em tempo polinomial*. Encontrar a resposta pode ser difícil, mas, uma vez encontrada, há um teste fácil de sua validade.

A expressão "não determinístico" é usada aqui pelo fato de ser possível solucionar um problema NP fazendo uma adivinhação inspirada. Tendo feito isso, você pode confirmar que ele está realmente correto (ou não). Por exemplo, se o problema é fatorar o número 11.111.111.111, você pode chutar que um fator seja o número primo 21.649. Assim apresentado, não passa de um chute a esmo. Mas é fácil checar: basta dividir por ele e ver o que se obtém. O resultado é exatamente 513.239, sem resto. Então o chute estava certo. Se em vez disso eu tivesse chutado 21.647 – que também é primo –, então a divisão daria 513.286 mais resto 9.069. E aí o chute teria sido errado.

Dar um chute certo é basicamente um milagre nessas circunstâncias, ou então existe algum truque (eu calculei os fatores de 11.111.111.111 antes de "chutar"). Mas na verdade é isso que queremos. Se não fosse algo milagroso, seria possível transformar um algoritmo classe NP em algoritmo classe P simplesmente dando montes e montes de chutes até um deles dar certo. Meu exemplo sugere por que isso não funciona: você precisa chutar demais. De fato, tudo que estamos fazendo aqui é uma "divisão por tentativas" por todos os fatores primos, até que um dê certo. Sabemos do Capítulo 2 que este é um meio desanimador de achar fatores.

A classe NP exclui exemplos bobos, tais como a minha lista muito longa. Se alguém adivinha uma lista de todos os dígitos binários de comprimento *n*, isso não só leva um tempo exponencial para imprimir, como também leva um tempo exponencial para ler, de modo que leva ainda mais tempo para se checar se está correta. Seria uma tarefa de leitura de provas

verdadeiramente terrível. A classe P decididamente está contida na classe NP. Se você puder achar a resposta em tempo polinomial, com garantia de estar correta, então a checagem já foi feita. Logo, a checagem automaticamente não exige nada pior do que um tempo polinomial. Se alguém lhe apresentasse a suposta resposta, você poderia processar o algoritmo inteiro novamente. A checagem é essa.

Agora podemos enunciar o problema do milênio. NP é maior que P ou são iguais? Resumindo: P é igual a NP?

Se a resposta for "sim" então seria possível encontrar algoritmos rápidos, eficientes, para programar horários de voos em linhas aéreas, otimizar a produção de uma fábrica ou executar um milhão de outras tarefas práticas importantes. Se a resposta for "não", teremos uma garantia rígida de que todos os problemas aparentemente difíceis são realmente difíceis, então poderemos parar de perder tempo tentando encontrar algoritmos rápidos para eles. De um modo ou de outro, saímos ganhando. Aborrecimento é não saber para que lado a coisa vai!

A vida seria muito mais simples para os matemáticos se a resposta fosse "sim", então o pessimista que habita em todo ser humano imediatamente desconfia de que a vida não será tão fácil assim, e a resposta provável é "não". Caso contrário, estaríamos todos ganhando um almoço grátis, que não merecemos e nada fizemos por merecer. Desconfio que a maioria dos matemáticos na realidade preferiria que a resposta fosse "não", porque isso os manteria ocupados até o fim da civilização. Os matemáticos se autoafirmam ao resolver problemas difíceis. Qualquer que seja a razão, a maioria dos matemáticos e cientistas da computação espera que a resposta à pergunta "P é igual a NP?" seja "não". Dificilmente alguém espera que seja "sim".

Há duas outras possibilidades. Pode ser possível provar que P é igual a NP sem efetivamente achar um algoritmo de tempo polinomial para qualquer problema NP específico. Os matemáticos têm o hábito de fornecer provas de existência que não são construtivas; eles mostram que algo existe, mas não dizem o que é. Exemplos incluem testes de primalidade, que alegremente nos informam que um número é primo sem apresentar qualquer fator específico, ou teoremas em teoria dos números afirmando que as soluções de alguma equação diofantina são inferiores a algum limite sem fornecer qualquer limite específico. O algoritmo de tempo polinomial poderia ser tão complicado que seria impossível escrevê-lo. Então o pessimismo natural relativo a almoços gratuitos estaria justificado, mesmo

que a resposta acabasse se revelando afirmativa.

De maneira mais drástica, alguns pesquisadores especulam que a questão pode ser indecidível dentro do corrente arcabouço lógico formal da matemática. Se assim for, não se pode provar nem "sim" nem "não". Não porque sejamos estúpidos demais para achar a prova, e sim porque a prova não existe. Essa opinião tornou-se visível em 1931 quando Kurt Gödel soltou o gato da indecidibilidade entre os pombos filosóficos que infestavam as fundações da matemática, provando que algumas afirmações em aritmética são indecidíveis. Em 1936, Alan Turing achou um problema indecidível mais simples, o da interrupção para as máquinas de Turing. Dado um algoritmo, existe sempre uma prova que a faça parar, ou uma prova que a faça continuar para sempre? A surpreendente resposta de Turing foi "não". Para alguns algoritmos não existe prova para nenhuma das duas coisas. O problema P/NP poderia, talvez, ser desse tipo. Isso explicaria por que ninguém consegue nem prová-lo nem refutá-lo. Mas ninguém consegue tampouco provar ou refutar que o problema P/NP é indecidível. Talvez sua indecidibilidade seja indecidível…

A MANEIRA MAIS DIRETA de abordar o problema P/NP seria escolher alguma questão conhecida como sendo da classe NP, assumir que exista um algoritmo de tempo polinomial para resolvê-la e, de algum modo, derivar uma contradição. Por algum tempo, as pessoas tentaram essa técnica numa variedade de problemas, mas em 1971 Stephen Cook percebeu que a escolha do problema geralmente não faz diferença. Existe uma sensação de que todos os problemas desse tipo – tirando ou adicionando alguns detalhes técnicos – sustentam-se ou desabam juntos. Cook introduziu a noção de um problema NP-completo. Trata-se de um problema NP específico, com a propriedade de que se existir um algoritmo classe P para resolvê-lo, então *qualquer* problema NP pode ser resolvido usando um algoritmo classe P.

Cook encontrou diversos problemas NP-completos, inclusive SAT, o problema da satisfatibilidade booleana, que indaga se uma dada expressão lógica pode se tornar verdadeira escolhendo a veracidade ou falsidade de suas variáveis de modo conveniente. Ele obteve também um resultado mais profundo: um problema mais restrito, 3-SAT, também é NP-completo. Aqui a fórmula lógica é uma que pode ser escrita na forma "*A* ou *B* ou *C* ou … ou *Z*", onde cada *A*, *B*, *C*, … *Z* é uma fórmula lógica envolvendo apenas três

variáveis. Não necessariamente as mesmas três variáveis a cada vez, apresso-me em acrescentar. A maioria das provas de que um dado problema é NP-completo recorre ao teorema de Cook sobre 3-SAT.

A definição de Cook implica que todos os problemas NP-completos estão no mesmo patamar. Provar que um deles é classe P provaria que todos são classe P. Este resultado deixa aberta uma possibilidade tática: alguns problemas NP-completos podem ser mais fáceis de trabalhar do que outros. Mas estrategicamente sugere que você pode muito bem pegar um problema NP-completo específico e trabalhar com ele. Problemas NP-completos sustentam-se ou caem juntos porque um problema NP-completo pode simular qualquer outro problema NP. Qualquer problema NP pode ser convertido num caso especial do NP-completo "codificando-o", utilizando um código que pode ser implantado em tempo polinomial.

Para ter um gostinho desse procedimento, considere um problema NP-completo típico: achar um ciclo hamiltoniano num grafo. Ou seja, especificar um trajeto fechado ao longo das arestas de um grafo que visite todo vértice (ponto) exatamente uma vez. Fechado significa que o trajeto retorna ao ponto de partida. O tamanho dos dados de entrada é o número de arestas, que é menor ou igual ao quadrado do número de vértices, pois cada aresta une dois vértices. (Assumimos que no máximo uma aresta una um dado par de vértices.) Não se conhece nenhum algoritmo classe P para solucionar este problema, mas suponhamos, hipoteticamente, que houvesse um. Agora vamos escolher algum outro problema, e chamá-lo de problema X. Suponhamos que o problema X possa ser reformulado em termos de encontrar um trajeto em algum grafo associado ao problema X. Se o método de traduzir os dados do problema X em dados sobre o grafo, e vice-versa, puder ser executado em tempo polinomial, então automaticamente obtemos um algoritmo classe P para o problema X, da seguinte maneira:

1. Traduzir o problema X em busca de um ciclo hamiltoniano num grafo correlato, o que pode ser feito em tempo polinomial.
2. Encontrar tal ciclo em tempo polinomial usando o algoritmo hipotético para o problema do grafo.
3. Traduzir de volta o resultante ciclo hamiltoniano para uma solução do problema X, o que novamente pode ser feito em tempo polinomial.

Uma vez que três passos de tempo polinomial combinados processam-se em tempo polinomial, este algoritmo é classe P.

Para mostrar como isso funciona, vou considerar uma versão menos ambiciosa do problema do ciclo hamiltoniano no qual não se exige que o trajeto seja fechado. Este é chamado problema do caminho hamiltoniano. Um grafo pode possuir um caminho hamiltoniano sem possuir um ciclo: a Figura 42 (esquerda) é um exemplo. Assim, uma solução para o problema do ciclo hamiltoniano pode não solucionar o problema do caminho hamiltoniano. No entanto, podemos converter o problema do caminho hamiltoniano num problema de ciclo hamiltoniano em um grafo correlacionado, mas diferente. Este é obtido adicionando-se um vértice extra. Ligado a cada vértice do grafo original como se vê na Figura 42 (direita). Qualquer ciclo hamiltoniano no novo grafo pode ser convertido num caminho hamiltoniano no grafo original: basta omitir o novo vértice e as duas arestas do ciclo que o encontram. Inversamente, qualquer caminho hamiltoniano no grafo original produz um ciclo hamiltoniano no novo grafo: basta ligar duas extremidades do caminho hamiltoniano ao novo ponto. Esta "codificação" do problema do trajeto como problema do ciclo introduz apenas um vértice novo e uma nova aresta por ponto original. Assim, o procedimento, e seu inverso, processam-se em tempo polinomial.

É claro que tudo que fiz aqui é codificar um problema específico como problema de ciclo hamiltoniano. Para provar que o problema do ciclo hamiltoniano é NP-completo, precisamos fazer o mesmo para qualquer problema NP. Isso pode ser feito: a primeira prova foi encontrada por Richard Karp em 1972, num famoso artigo que provou que 21 problemas diferentes são NP-completos.[3]

FIGURA 42 *Esquerda*: Grafo com um caminho hamiltoniano (linha cheia), mas sem ciclo hamiltoniano. *Direita*: Acrescenta-se um ponto extra (cinza) e quatro novas arestas para converter o caminho hamiltoniano em ciclo hamiltoniano (linha cheia). As duas arestas cinzentas não estão no ciclo mas são necessárias para a construção de um grafo maior.

O problema do caixeiro-viajante é "quase" NP-completo, mas há uma

questão técnica: ele não é conhecido como sendo NP. São conhecidos mais de trezentos problemas específicos NP-completos, em áreas da matemática que incluem lógica, grafos, combinatória e otimização. Provar que qualquer um deles pode ou não ser resolvido em tempo polinomial provaria o mesmo para cada um deles. Apesar dessa constrangedora opulência, o problema P/NP permanece em aberto. E não me surpreenderia que ainda estivesse daqui a cem anos.

12. Pensamento fluido

A equação de Navier-Stokes

CINCO DOS PROBLEMAS DO MILÊNIO, inclusive os três discutidos até aqui, vêm da matemática pura, embora o problema P/NP também seja fundamental em ciência da computação. Os outros dois vêm da matemática clássica aplicada e da moderna física matemática. O problema da matemática aplicada surge de uma equação-padrão para fluxo de fluidos, a equação de Navier-Stokes, que recebeu este nome em virtude do engenheiro e físico francês Claude-Louis Navier e do matemático e físico irlandês George Stokes. Sua equação é uma equação diferencial parcial, o que significa que envolve a taxa de variação do padrão de fluxo tanto no espaço como no tempo. A maioria das grandes equações da física e da matemática clássica aplicada também são equações diferenciais parciais – acabamos de conhecer uma, a de Laplace –, e aquelas que não o são, são equações diferenciais comuns, envolvendo apenas a taxa de variação em relação ao tempo.

No Capítulo 8 vimos como o movimento do sistema solar é determinado pela lei da gravitação e do movimento de Newton. Elas relacionam as acelerações do Sol, da Lua e dos planetas com as forças gravitacionais que estão atuando. Aceleração é a taxa de variação da velocidade em relação ao tempo, e velocidade é a taxa de variação da posição em relação ao tempo. Então esta é uma equação diferencial comum. Como vimos, resolver tais equações pode ser bastante difícil. Resolver equações diferenciais parciais geralmente é mais difícil ainda.

Para propósitos práticos, as equações do sistema solar podem ser resolvidas numericamente usando-se computadores. Ainda assim é difícil, mas atualmente existem bons métodos. O mesmo vale para aplicações práticas das equações de Navier-Stokes. As técnicas empregadas são conhecidas como dinâmica dos fluidos computacional e têm vasta gama de aplicações importantes: projetos de aviões, aerodinâmica de carros, até mesmo problemas médicos como a corrente sanguínea no corpo humano.

O problema-prêmio do milênio não pede aos matemáticos que encontrem soluções explícitas para a equação de Navier-Stokes, porque isso é essencialmente impossível. Tampouco trata de métodos numéricos

para resolver as equações, por mais importantes que estes sejam. Em vez disso, pede uma prova de uma propriedade teórica básica: a *existência* de soluções. Dado o estado de um fluido em determinado instante no tempo – o padrão em que está se movendo – existirá uma solução para a equação de Navier-Stokes, válida para todo o tempo futuro, a partir do referido estado? A intuição física sugere que a resposta deve seguramente ser "sim", porque a equação é um modelo muito preciso da física de fluidos reais. Entretanto, a questão matemática da existência não é tão nítida, e essa propriedade básica da equação jamais foi provada. Pode ser que nem mesmo seja verdade.

A EQUAÇÃO DE NAVIER-STOKES descreve como o padrão de velocidade dos fluidos varia com o tempo, em circunstâncias dadas. A equação é muitas vezes mencionada usando-se o plural, equações de Navier-Stokes, mas dá no mesmo. O plural reflete a visão clássica: no espaço tridimensional, a velocidade tem três componentes, e classicamente cada componente contribui com uma equação, sendo três no total. Na visão moderna, há uma equação para o *vetor* velocidade (uma grandeza com tamanho, direção e sentido), mas essa equação pode ser aplicada a cada uma das três componentes da velocidade. O site do Instituto Clay usa a terminologia clássica, mas aqui seguirei a prática moderna. Estou mencionando isso para evitar possível confusão.

A equação data de 1822, quando Navier escreveu uma equação diferencial parcial para o escoamento de um fluido viscoso – pegajoso. As contribuições de Stokes ocorreram em 1842 e 1843. Euler havia escrito uma equação diferencial parcial para um fluido com viscosidade zero – não pegajoso – em 1757. Embora essa equação continue útil, a maioria dos fluidos reais, inclusive água e ar, são viscosos; assim, Navier e Stokes modificaram a equação Euler de modo a levar em conta a viscosidade. Os dois cientistas deduziram essencialmente a mesma equação de forma independente, por isso ela recebe o nome de ambos. Navier cometeu alguns erros matemáticos, mas chegou à resposta certa; Stokes acertou na matemática, e é por isso que sabemos que a resposta de Navier está correta, apesar do seu engano. Em sua forma mais geral, a equação aplica-se a fluidos compressíveis como o ar. Contudo, há um caso especial importante no qual presume-se que o fluido seja incompressível. Tal modelo aplica-se a fluidos como a água, que realmente se comprimem sob forças muito grandes, mas apenas ligeiramente.

Há duas maneiras de se montar a descrição matemática do escoamento de fluidos: pode-se descrever a trajetória que cada partícula de fluido percorre com o passar do tempo, ou pode-se descrever a velocidade do fluxo em cada ponto no espaço e cada instante no tempo. As duas descrições estão correlacionadas: dada uma, pode-se – com esforço – deduzir a outra. Euler, Navier e Stokes, todos eles usaram o segundo ponto de vista, porque este conduz a uma equação muito mais tratável matematicamente. Então, suas equações referem-se a campo de velocidade do fluido. Em cada instante de tempo fixo, o campo de velocidade especifica o valor, a direção e o sentido da velocidade de cada partícula de fluido. À medida que o tempo varia, essa descrição pode mudar. É por isso que nessa equação ocorrem taxas de variação tanto no espaço como no tempo.

A equação de Navier-Stokes tem um excelente pedigree em física. Ela se baseia nas leis do movimento de Newton, aplicadas a cada partícula minúscula (pequena região) do fluido, e exprime, nesse contexto, a lei da conservação da quantidade de movimento. Cada partícula se move porque forças atuam sobre ela, e a lei do movimento de Newton afirma que a aceleração da partícula é proporcional à força. As principais forças são atrito, causado pela viscosidade, e pressão. Há também forças geradas pela aceleração da partícula. A equação segue a prática clássica, e trata o fluido como um continuum infinitamente divisível. Em particular, ela ignora a estrutura atômica discreta do fluido em escalas muito pequenas.

As equações em si têm pouco valor: você precisa ser capaz de resolvê-las. Para a equação de Navier-Stokes, isso significa calcular o campo de velocidade: o valor, a direção e o sentido da velocidade do fluido em cada ponto no espaço e em cada instante no tempo. A equação fornece restrições para essas grandezas, mas não as prescreve diretamente. Em vez disso, temos de aplicar a equação para relacionar as velocidades futuras com as atuais. Equações diferenciais parciais como a de Navier-Stokes têm muitas soluções distintas; na verdade, infinitas. Isso não é surpresa: fluidos podem correr de muitas maneiras diferentes – o fluxo sobre a superfície de um carro difere do fluxo sobre as asas de um avião. Estes são dois modos de selecionar um fluxo particular entre a infinidade de possibilidades: condições de contorno e condições de fronteira.

Condições de contorno especificam o campo de velocidade em um determinado instante de referência, geralmente tomado como instante zero. A ideia física é que uma vez que se conheça o campo de velocidade neste instante, a equação de Navier-Stokes determina o campo após um

intervalo de tempo muito curto, de forma única. Se você começa dando um impulso no fluido, ele segue adiante obedecendo às leis da física. As condições são mais úteis na maioria das aplicações, porque é difícil estabelecer condições iniciais num fluido real, e em todo caso estas não são inteiramente apropriadas a aplicações como o projeto de um carro. O que importa aí é o formato do carro. Fluidos viscosos aderem a superfícies. Matematicamente essa característica é modelada especificando-se a velocidade sobre essas superfícies, que forma a fronteira da região ocupada pelo fluido, que é onde a equação é válida. Por exemplo, podemos querer que a velocidade seja zero na fronteira, ou qualquer outra condição que melhor modele a realidade.

Mesmo quando as condições de contorno ou de fronteira são especificadas, é altamente incomum conseguir anotar uma fórmula explícita para o campo de velocidade, pois a equação de Navier-Stokes é não linear. A soma de duas soluções normalmente não é uma solução. Existe um motivo para que o problema dos três corpos no Capítulo 8 seja tão difícil – embora não seja o único, porque o problema dos dois corpos também é não linear, apesar de ter uma solução explícita.

Para propósitos práticos, podemos resolver a equação de Navier-Stokes em um computador, representando o campo de velocidade como uma lista de números. Essa lista pode ser transformada em gráficos elegantes e usada para calcular grandezas do interesse de engenheiros, tais como as pressões sobre as asas do avião. Como os computadores não conseguem manipular listas infinitas de números, nem lidar com números numa precisão infinita, temos de substituir o fluxo real por uma aproximação discreta, ou seja, uma lista de números que mostre o fluxo numa quantidade finita de locais e instantes. A grande questão é assegurar que a aproximação seja boa o bastante.

A abordagem comum é dividir o espaço numa grande quantidade de pequenas regiões de modo a formar uma grade computacional. A velocidade é calculada apenas para pontos nos cantos da grade. Esta pode ser simplesmente um conjunto de quadrados (ou cubos, em três dimensões), como um tabuleiro de xadrez, mas para carros e aviões precisa ser mais complicada, com regiões perto da fronteira, para capturar detalhes mais finos do fluxo. A grade pode ser dinâmica, mudando de forma com o passar do tempo. O tempo geralmente é assumido progredindo em passos, que podem ter todos o mesmo tamanho, ou variar de acordo com o estado predominante do cálculo.

A base da maioria dos métodos numéricos é a forma como “a taxa de variação” é definida no cálculo. Suponha que um objeto se mova de um local para outro num intervalo de tempo muito curto. Então a taxa de variação da posição – a velocidade – é a variação da posição dividida pelo tempo que levou, com um pequeno erro que desaparece à medida que o intervalo de tempo fica menor. Assim, podemos fazer a aproximação da taxa de variação, que é o que entra na equação de Navier-Stokes, mediante esta razão entre variação espacial e variação temporal. Efetivamente, a equação agora nos diz como empurrar um estado inicial conhecido – uma especificada lista de velocidades – um passo adiante no futuro. Aí precisamos repetir os cálculos muitas vezes, e ver o que acontece mais adiante no futuro. Há um meio similar de aproximar soluções quando a que queremos é determinada por condições de fronteira. Há também muitos meios sofisticados de se conseguir o mesmo resultado com maior precisão.

Quanto mais fina for a divisão da grade computacional, e menores forem os intervalos de tempo, mais acurada torna-se a aproximação. Contudo, a computação também leva mais tempo. Assim, existe um meio-termo de compromisso entre precisão e rapidez. Em termos gerais, uma resposta aproximada obtida por computador é propensa a ser aceitável, contanto que o fluxo não tenha características significativas que sejam menores do que o tamanho da grade. Há dois tipos principais de fluxo: laminar e turbulento. No fluxo laminar, o padrão de movimento é tranquilo, e camadas de fluido deslizam suavemente uma pela outra. Aqui uma grade suficientemente pequena deve servir. O fluxo turbulento é mais violento e espumoso, e o fluido se mistura de maneiras extremamente complexas. Em tais circunstâncias, uma grade discreta, por mais fina que seja, pode facilmente causar problemas.

Uma das características da turbulência é a ocorrência de vórtices, como pequenos redemoinhos, e que podem de fato ser muito pequenos. Uma imagem habitual da turbulência consiste numa cascata de vórtices cada vez menores. A maior parte do pequeno detalhe é menor do que qualquer grade prática. Para contornar esta dificuldade os engenheiros frequentemente recorrem a modelos estatísticos em questões sobre fluxo turbulento. Outra preocupação é que o modelo físico de um continuum pode ser inapropriado para o fluxo turbulento, porque os vórtices podem se reduzir a proporções atômicas. Entretanto, comparações entre cálculos numéricos e experimentos mostram que a equação de Navier-Stokes é um modelo bastante acurado e realista – tão bom que muitas aplicações em

engenharia agora se assentam apenas em dinâmica de fluidos computacional, que é barata, em vez de realizar experimentos com modelos em escala em túneis de vento, que são caros. Contudo, verificações experimentais como essas ainda são usadas quando a segurança humana é vital – por exemplo, quando se projetam aviões.

Na verdade, a equação de Navier-Stokes é tão precisa que parece se aplicar mesmo quando a física sugere que há uma razoável chance de falhar: no fluxo turbulento. Pelo menos, este é o caso se ela puder ser resolvida com suficiente precisão. O principal problema é prático: métodos numéricos para resolver a equação tomam enormes quantidades de tempo de computador quando o fluxo se torna turbulento. E sempre perdem algum detalhe da estrutura em pequena escala.

OS MATEMÁTICOS SEMPRE se sentem desconfortáveis quando a principal informação que possuem sobre um problema baseia-se em algum tipo de aproximação. O prêmio do milênio para a equação de Navier-Stokes ataca um dos temas teóricos fundamentais. Sua solução reforçaria a sensação visceral de que métodos numéricos geralmente funcionam muito bem. Há uma distinção sutil entre as aproximações usadas por um computador, que fazem pequenas mudanças na equação, e a precisão da resposta, que trata de pequenas mudanças na solução. Será que uma resposta exata para uma pergunta aproximada é a mesma coisa que uma resposta aproximada para uma pergunta exata? Às vezes a resposta é "não". O escoamento exato de um fluido com viscosidade muito pequena, por exemplo, várias vezes difere de um escoamento aproximado para um fluido com viscosidade zero.

Um passo adiante para a compreensão desses assuntos é tão simples que pode ser facilmente esquecido: provar que existe uma solução exata. Deve haver alguma coisa *em relação à qual* os cálculos do computador devem se aproximar. É esta observação que motiva o prêmio do milênio para a equação de Navier-Stokes. Sua descrição oficial no site do Instituto Clay consiste em quatro problemas. Solucionar qualquer um deles é suficiente para ganhar o prêmio. Em todos os quatro, o fluido é presumido incompressível. São eles:

1. *Existência e regularidade de soluções em três dimensões*. Aqui admite-se que o fluido preencha todo o espaço infinito. Dado qualquer campo de velocidade regular inicial, provar que existe uma solução regular para a

equação para todos os tempos positivos, coincidindo com o campo inicial especificado.

2. *Existência e regularidade de soluções no toro plano tridimensional*. A mesma questão, porém agora admitindo que o espaço seja um toro plano – uma caixa retangular com faces opostas coincidindo. Essa versão evita problemas potenciais causados pelo domínio infinito assumido na primeira versão, que não combina com a realidade e pode causar mau comportamento por razões tolas.
3. *Colapso de soluções em três dimensões.* Provar que (1) está errado. Ou seja, encontrar um campo inicial para o qual a solução regular não existe para todos os tempos positivos, e provar essa afirmação.
4. *Colapso de soluções no toro tridimensional*. Provar que (2) está errado.

Os mesmos problemas permanecem em aberto para a equação de Euler, que é a mesma que a equação de Navier-Stokes mas presume ausência de viscosidade, entretanto não existe prêmio oferecido para a equação de Euler.

A grande dificuldade aqui é que o fluxo em consideração é tridimensional. Há uma equação análoga para fluidos correndo no plano. Fisicamente, este representa uma fina camada de fluido entre duas placas planas, presumindo não causar atrito, ou então um padrão de fluxo em três dimensões no qual o fluido move-se exatamente da mesma maneira ao longo de um sistema de planos paralelos. Em 1969, a matemática russa Olga Alexandrovna Ladyzhenskaya provou que (1) e (2) são verdadeiros, enquanto (3) e (4) são falsos, para a equação de Navier-Stokes bidimensional e para a equação de Euler bidimensional.

Talvez surpreendentemente, a prova é mais difícil para a equação de Euler, mesmo que seja uma equação mais simples que a de Navier-Stokes, pois omite termos envolvendo viscosidade. A razão é instrutiva: a viscosidade "amortece" o mau comportamento na solução, que potencialmente poderia levar a algum tipo de singularidade impedindo que a solução existisse para todo tempo. Se o termo da viscosidade está ausente, este amortecimento não ocorre, e isso se manifesta como problemas matemáticos na prova existente.

Ladyzhenskaya fez outras contribuições vitais para a nossa compreensão da equação de Navier-Stokes, provando não só que existem soluções, mas também que certos esquemas de dinâmica de fluidos

computacional geram aproximações tão precisas quanto desejamos.

Os PROBLEMAS DO PRÊMIO do milênio referem-se a um fluxo incompressível, porque fluxos compressíveis são conhecidos pelo seu mau comportamento. As equações para um avião, por exemplo, enfrentam todo tipo de problemas se o avião viaja mais rápido do que o som. É a famosa "barreira do som", que preocupou engenheiros tentando projetar caças a jato supersônicos, e o problema está relacionado com a compressibilidade do ar. Se um corpo se move através de um fluido incompressível, ele empurra as partículas do fluido tirando-as da frente do caminho, do mesmo modo que passar por dentro de um túnel cheio de rolamentos. Se as partículas se acumulam, retardam o movimento do corpo. Mas num fluido compressível, onde há um limite para a velocidade em que as ondas podem viajar – a velocidade do som –, isso não acontece. Em velocidades supersônicas, em vez de serem empurradas para fora do caminho, o ar se acumula na frente do avião, e sua densidade aumenta então de forma ilimitada. O resultado é uma onda de choque. Matematicamente, essa é uma descontinuidade na pressão do ar, que de repente salta de um valor para outro diferente através da onda de choque. Fisicamente, o resultado é o estouro sônico: uma forte explosão sonora. Se não compreendida e levada em conta, uma onda de choque pode danificar o avião, então os engenheiros tinham razão em se preocupar. Mas a velocidade do som não é realmente uma barreira, apenas um obstáculo. A presença de ondas de choque significa que as equações de Navier-Stokes compressíveis não precisam ter soluções regulares para todo tempo, mesmo em duas dimensões. Logo, a resposta nesse caso já é conhecida, e é negativa.

A matemática das ondas de choque é uma área substancial dentro das equações diferenciais parciais, apesar do colapso das soluções. Embora a equação de Navier-Stokes sozinha não seja um bom modelo físico para fluidos compressíveis, é possível modificar o modelo matemático acrescentando condições extras às equações, que levam em conta descontinuidades das ondas de choque. Estas, porém, não ocorrem no escoamento de um fluido incompressível, então é pelo menos concebível que nesse contexto as soluções devam existir para todo tempo, não importa quão complicado possa ser o fluxo inicial, contanto que seja regular.

Alguns resultados positivos são conhecidos para a equação de Navier-Stokes tridimensional. Se o padrão inicial de fluxo envolve velocidades

suficientemente pequenas, de modo que o escoamento seja muito lento, então (1) e (2) são ambos verdadeiros. Mesmo se as velocidades forem grandes, (1) e (2) são verdadeiros para algum intervalo de tempo diferente de zero. Pode não existir uma solução válida para todos os tempos futuros, mas há uma quantidade finita de tempo ao longo do qual uma solução existe. Pode parecer que é possível repetir esse processo empurrando a solução adiante no tempo em intervalos pequenos, e aí usando o resultado final como uma nova condição inicial. O problema nessa linha de raciocínio é que os intervalos de tempo podem encolher tão rapidamente que infinitos passos desse tipo levem um tempo finito. Por exemplo, se cada passo sucessivo levar metade do tempo que o anterior, e o primeiro passo leva, digamos, um minuto, então o processo todo termina num tempo de $1 + {}^{1}/_{2} + {}^{1}/_{4} + {}^{1}/_{8} + \ldots$, que perfaz 2. Se a solução deixa de existir – no presente apenas uma premissa puramente hipotética, mas que ainda assim podemos contemplar – então diz-se que a referida solução *explode* o tempo. O tempo que leva para isso acontecer é o "tempo de explosão".

Os quatro problemas indagam se as soluções podem, efetivamente, explodir. Se não puderem, (1) e (2) são verdadeiros; se puderem, (3) e (4) o são. Talvez soluções possam explodir num domínio infinito, mas não num domínio finito. Aliás, se a resposta a (1) for "sim", então a resposta a (2) também será, porque podemos interpretar qualquer padrão de fluxo num toro plano como um padrão de fluxo espacialmente periódico na totalidade do espaço infinito. A ideia é preencher o espaço com cópias da caixa retangular envolvida e copiar o mesmo padrão de fluxo em cada uma delas. As regras de colagem para o toro asseguram que o fluxo permaneça regular ao cruzar essas interfaces planas. Da mesma forma, se a resposta a (4) for "sim", a resposta a (3) também será, pelo mesmo motivo. Apenas tornamos o estado inicial espacialmente periódico. Mas por tudo que sabemos atualmente, a resposta a (2) pode ser "sim" mas a resposta a (1) pode ser "não".

Com efeito, conhecemos um fato surpreendente a respeito das explosões. Se houver uma solução com um tempo de explosão finito, então a velocidade máxima do fluido, em todos os pontos do espaço, precisa se tornar tão grande quanto se queira. Isso poderia ocorrer, por exemplo, se se formar um jato de fluidos e a velocidade do jato aumentar tão rapidamente que divergirá ao infinito após ter se passado uma quantidade de tempo finita.

ESSAS OBJEÇÕES não são puramente hipotéticas. Há precedentes para esse tipo de comportamento singular em outras equações da física matemática clássica. Um exemplo notável ocorre em mecânica celeste. Em 1988, Zhihong Xia provou que existe uma configuração inicial de cinco massas puntiformes no espaço tridimensional, obedecendo à lei da gravitação de Newton, para a qual quatro partículas desaparecem no infinito após um período de tempo finito – uma forma de explosão – e a quinta passa por oscilações cada vez mais violentas. Antes disso, Joseph Gerver já indicara que cinco corpos num plano podem todos desaparecer no infinito num tempo finito, mas foi incapaz de completar a prova para o cenário que divisou. Em 1989, ele provou que esse tipo de escape decididamente pode ocorrer num plano se o número de corpos for grande o suficiente.

É extraordinário que esse comportamento seja possível, considerando que tais sistemas obedecem à lei da conservação da energia. É certo que, se todos os corpos estiverem se movendo de forma arbitrariamente rápida, a energia cinética total não deve aumentar? A resposta é que há também uma redução de energia potencial, e para uma partícula puntiforme, a energia potencial gravitacional total é infinita. Os corpos precisam conservar também o momento angular, mas podem fazê-lo se alguns deles moverem-se cada vez mais depressa em círculos de raios cada vez menores.

O ponto físico envolvido é o famoso efeito estilingue, usado rotineiramente para despachar sondas espaciais a mundos distantes no sistema solar. Um bom exemplo é a sonda Galileo, da Nasa, cuja missão era viajar a Júpiter para estudar o planeta gigante e seus muitos satélites. Foi lançada em 1989 e chegou a Júpiter em 1995. Uma das razões para ter levado tanto tempo foi que sua rota era distintamente indireta. Embora a órbita de Júpiter fique fora da órbita da Terra, Galileo começou dirigindo-se para dentro, rumo a Vênus. Chegou próximo a esse planeta, retornou e passou voando pela Terra rumando para o espaço exterior para examinar o asteroide 951 Gaspra. Depois voltou rumo à Terra, deu *mais* uma volta pelo nosso planeta e finalmente rumou para Júpiter. No caminho aproximou-se de outro asteroide, Ida, descobrindo que ele tinha sua própria e diminuta lua, um novo asteroide chamado Dactyl.

Por que uma trajetória tão complicada? Galileo ganhava energia, portanto velocidade, a cada nova aproximação. Imagine uma sonda espacial rumando para um planeta que se aproxima, não em rota de colisão, mas chegando suficientemente perto da superfície, dando uma volta pela parte de trás do planeta e sendo lançada novamente no espaço. Quando a sonda

passa por trás do planeta, os dois se atraem mutuamente. Na verdade, estão sendo atraídos um pelo outro o tempo todo, mas nessa fase a força de atração está no seu máximo e portanto tem o maior efeito. A gravidade do planeta dá à sonda um impulso de velocidade. A energia precisa ser conservada, então, em compensação, a sonda reduz ligeiramente a velocidade do planeta em sua órbita em torno do Sol. Como a sonda tem uma massa muito pequena e o planeta uma massa muito grande, o efeito sobre o planeta é desprezível. O efeito sobre a sonda não é: ela pode acelerar drasticamente.

A Galileo chegou a menos de 16 mil quilômetros da superfície de Vênus, e ganhou 2,23 quilômetros por segundo de velocidade. Então passou a 960 quilômetros da Terra, e mais uma vez a trezentos quilômetros, adicionando mais 3,7 quilômetros por segundo. Essas manobras foram essenciais para levá-la a Júpiter, porque seus foguetes não tinham potência suficiente para levá-la diretamente até lá. O plano original tinha sido fazer exatamente isso, usando o foguete Centaur-G, à base de combustível hidrogênio líquido. Mas o desastre no qual o ônibus espacial Challenger explodiu logo após o lançamento levou esse plano a ser abandonado, porque o Centaur-G foi proibido. Assim, a Galileo precisou usar um foguete mais fraco à base de combustível sólido. A missão foi um enorme sucesso, e a recompensa científica incluiu observar a colisão entre o cometa Shoemaker-Levy 9 e Júpiter em 1994, enquanto a sonda ainda estava a caminho de Júpiter.

O cenário de Xia também faz uso do efeito estilingue. Quatro planetas de igual massa formam dois pares próximos, girando em torno de seus centros de massa comuns em dois planos paralelos.[1] Essas duas raquetes de dois corpos jogam tênis celeste usando como bola um quinto corpo, mais leve, que oscila de um lado a outro entre elas em ângulos retos com os dois planos. O sistema é montado de modo que toda vez que a “bola de tênis” passa por um par de planetas, o efeito estilingue acelera a bola e empurra o par de planetas para fora ao longo da linha que une os dois pares, de modo que a quadra de tênis vai ficando cada vez mais comprida e os jogadores cada vez mais distantes entre si. Energia e quantidade de movimento são mantidas em equilíbrio porque os dois planetas em questão aproximam-se ligeiramente e giram cada vez mais rápido em torno de seu centro de massa. Com a configuração inicial adequada, os pares de planetas vão se afastando cada vez mais rápido, e sua velocidade aumenta tão depressa que chegam ao infinito após uma quantidade de tempo finita. Entretanto, a bola de tênis continua oscilando entre “as raquetes” cada vez mais depressa. Os

cenários de escape de Gerver também utilizam o efeito estilingue.

Será esse número de desaparecimento relevante para corpos celestes reais? Não se for tomado ao pé da letra. Ele se fundamenta nos corpos como massas puntiformes. Para muitos problemas em mecânica celeste, trata-se de uma aproximação coerente, mas não se os corpos chegam arbitrariamente muito perto um do outro. Se corpos de tamanho finito fizessem isso, acabariam colidindo. Efeitos relativistas impediriam os corpos de se mover mais depressa que a luz e mudariam a lei da gravitação. De qualquer modo, as condições de contorno e as premissas de que algumas massas são idênticas, seriam raras demais para ocorrer na prática. Não obstante, esses exemplos curiosos mostram que mesmo que as equações da mecânica celeste modelem bastante bem a realidade na maioria das circunstâncias, podem ter complicadas singularidades que impedem a existência das soluções para todo tempo. Recentemente também se descobriu que o efeito estilingue em sistemas de estrelas triplas, onde as estrelas orbitam uma em torno da outra em trajetórias complicadas, pode expelir uma das estrelas em alta velocidade. Portanto, pode haver por aí inúmeras estrelas órfãs, expulsas de seus sistemas pelas suas irmãs, vagando pela galáxia – ou mesmo no espaço intergaláctico – frias, solitárias, indesejadas e despercebidas.

QUANDO UMA EQUAÇÃO diferencial se comporta de uma forma tão estranha que suas soluções deixam de fazer sentido após algum período de tempo finito, dizemos que há uma singularidade. O trabalho acima a respeito do problema de muitos corpos é na realidade sobre vários tipos de singularidade. O problema do milênio sobre a equação de Navier-Stokes pergunta se podem ocorrer singularidades em problemas de valor inicial para um fluido ocupando a totalidade do espaço ou um toro plano. Se uma singularidade puder se formar num tempo finito, o resultado é provavelmente uma explosão, a menos que a singularidade se deslinde subsequentemente, o que parece improvável.

Há dois modos principais de se abordar essas questões. Podemos tentar provar que singularidades nunca ocorrem, ou tentarmos encontrar uma delas escolhendo condições de contorno convenientes. Soluções numéricas podem ajudar, de um modo ou de outro: podem sugerir características gerais úteis dos fluxos e podem fornecer fortes indícios da possível natureza de singularidades potenciais. No entanto, a potencial falta de

precisão nas soluções numéricas significa que qualquer indício desses deve ser tratado com cautela e justificado com mais rigor.

Tentativas de provar regularidade – ausência de singularidades – empregam uma variedade de métodos para ganhar controle sobre o fluxo. Esses métodos incluem complicadas estimativas de quão grandes ou pequenas podem se tornar certas variáveis básicas, ou então técnicas mais abstratas. Uma abordagem popular é por meio das soluções fracas, que não são exatamente fluxos, mas estruturas matemáticas mais gerais com algumas das propriedades de fluxos. Sabe-se, por exemplo, que o conjunto de singularidades de uma solução fraca da equação tridimensional de Navier-Stokes é sempre pequeno, num sentido técnico específico.

Muitos cenários diferentes que poderiam levar a singularidades têm sido investigados. O modelo-padrão de turbulência como cascata de vórtices sempre decrescentes remonta a Andrei Kolmogorov, em 1941. Ele sugeriu que, em escalas muito pequenas, todas as formas de turbulência têm aspecto muito similar. A proporção de vórtices de determinado tamanho, por exemplo, segue uma lei universal. Agora se sabe que à medida que os vórtices vão ficando menores, mudam de formato, tornando-se mais longos e finos, formando filamentos. A lei da conservação do momento angular implica que a vorticidade – quanto o vórtice está girando – deve aumentar. Isto é chamado espichamento de vórtice, e é um tipo de comportamento que pode provocar singularidade – por exemplo, se vórtices muito pequenos pudessem se tornar infinitamente longos num tempo finito e a vorticidade pudesse se tornar infinita em alguns pontos.

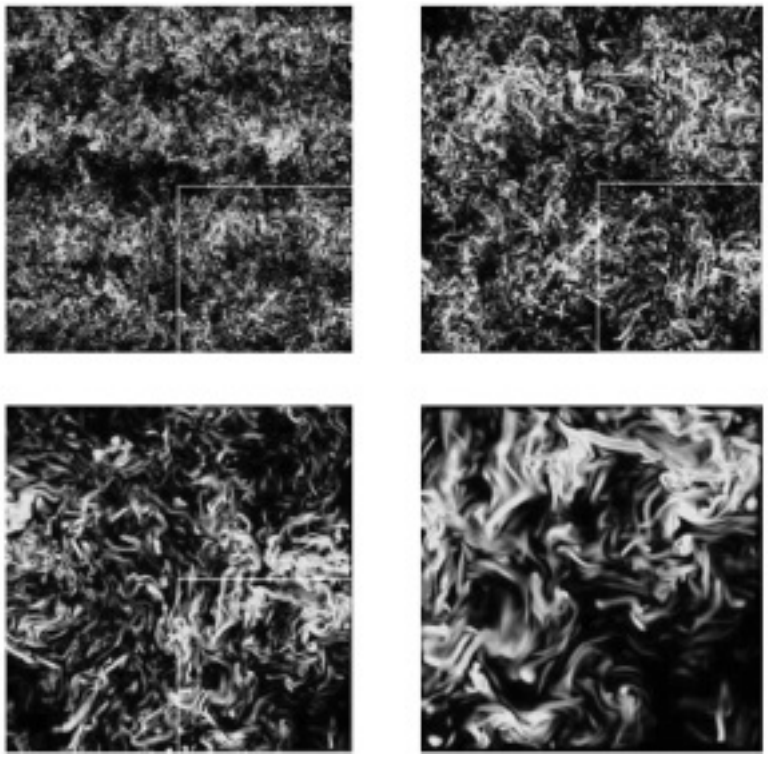


FIGURA 43 Ampliações de um fluxo turbulento, simulado com o sistema computadorizado Vapor.

A Figura 43 mostra aproximações em escalas muito pequenas de um fluxo turbulento, simulado por Pablo Mininni e seus colaboradores usando o Vapor – Visualization and Analysis for Ocean, Atmosphere, and Solar Research (Visualização e Análise para Pesquisa Oceânica, Atmosférica e Solar). As imagens mostram a intensidade da vorticidade: a rapidez com que o fluido está girando. Elas ilustram a formação de filamentos de vórtice: as estruturas longas e finas nas figuras, e mostram que elas podem se agrupar para formar padrões em escala maior. A configuração é capaz de executar simulações em grades cúbicas com mais de 3 bilhões de pontos.

Em seu artigo sobre este problema no site do Instituto Clay,[2] Charles Fefferman escreve:

Há muitos problemas e conjecturas fascinantes sobre o comportamento de soluções das equações de Euler e Navier-Stokes … Como nem sequer sabemos se essas soluções existem, nossa compreensão está num nível muito primitivo. Métodos padronizados de [equações diferenciais parciais] parecem inadequados para resolver o problema. Em vez disso, provavelmente necessitamos de algumas ideias novas, profundas.

A complexidade do fluxo em imagens como as da Figura 43 deixam

claras as dificuldades propensas a serem encontradas ao buscarmos tais ideias. Destemidos, os matemáticos estão cerrando fileiras, buscando princípios simples dentro de aparentes complexidades.

13. Enigma quântico

A hipótese do mass gap

A POUCOS QUILÔMETROS de Genebra há uma brusca dobra na fronteira entre a Suíça e a França. Na superfície, tudo o que se vê são estradinhas secundárias e pequenos vilarejos. Mas entre cinquenta e 175 metros abaixo, no subsolo, está o maior instrumento científico do planeta. É um gigantesco túnel circular com mais de oito quilômetros de diâmetro, ligado a um segundo túnel circular com cerca de um quarto desse tamanho. A maior parte está sob a França, mas duas seções estão na Suíça. Dentro dos túneis correm pares de tubos, que se juntam em quatro pontos.

É o Grande Colisor de Hádrons, que custou 7,5 bilhões de euros (cerca de 9 bilhões de dólares) e está sondando as fronteiras da física de partículas. O principal objetivo dos 10 mil cientistas de mais de cem países que colaboraram no equipamento era encontrar o bóson de Higgs – ou não encontrar, se fosse este o modo de o continuum se desfazer. Estavam à procura do bóson de Higgs para completar o Modelo Padrão da física de partículas, no qual tudo no universo é feito segundo dezessete diferentes partículas fundamentais. De acordo com a teoria, o bóson de Higgs é o que dá massa a todas as partículas.

Em dezembro de 2011, Atlas e CMS, duas divisões experimentais do Grande Colisor de Hádrons, descobriram de forma independente evidência experimental de um bóson de Higgs com massa de aproximadamente 125 GeV (gigaeletronvolts, unidade usada indistintamente para massa e energia em física de partículas, já que ambas são equivalentes). Em 4 de julho de 2012, o Cern, o laboratório europeu de física de partículas que opera o Grande Colisor de Hádrons, anunciou, para uma plateia lotada de cientistas e jornalistas de ciência, que o continuum se desfizera em favor do Higgs. Ambos os grupos haviam coletado grandes quantidades de dados adicionais, e a chance de os dados mostrarem uma flutuação aleatória, e não uma nova partícula com propriedades tipo Higgs, havia caído abaixo de um em 2 milhões. Este é o grau de confiança tradicionalmente exigido em física de partículas antes de estourar o champanhe.

Experimentos adicionais serão necessários para assegurar que a nova

partícula tenha todas as características que o bóson de Higgs teórico deve possuir. Por exemplo, a teoria prediz que o bóson de Higgs deve ter spin 0; na época do anúncio, a observação mostrava que era 0 ou 2. Há também a possibilidade de que "o" bóson de Higgs acabe se revelando composto de outras partículas, menores, ou que seja apenas o primeiro de uma nova família de partículas tipo Higgs. Logo, ou o modelo corrente de partículas fundamentais será cimentado no lugar, ou teremos nova informação que eventualmente nos conduzirá a uma teoria melhor.

O último dos sete problemas do milênio está intimamente relacionado com o Modelo Padrão e o bóson de Higgs. Trata-se de uma questão central na teoria quântica de campo, o arcabouço matemático no qual se estuda física de partículas. Chama-se hipótese da ausência de massa, e coloca um limite inferior específico para a possível massa de uma partícula fundamental. É um problema representativo, escolhido dentro de uma série de grandes questões não resolvidas nessa área profunda e muito nova da física matemática. Ele tem ligações que vão desde as fronteiras da matemática pura até a tão aguardada unificação das duas principais teorias físicas, a relatividade geral e a teoria quântica de campo.

EM MECÂNICA CLÁSSICA newtoniana, as grandezas físicas básicas são espaço, tempo e massa. O espaço é assumido como euclidiano tridimensional, o tempo é uma grandeza unidimensional independente do espaço, e a massa significa presença de matéria. As massas variam sua posição no espaço sob a influência de forças, e essa taxa de variação de posição é medida em relação ao tempo. A lei do movimento de Newton descreve como a aceleração de um corpo (a taxa de variação da sua velocidade, que por sua vez é a taxa de variação da posição) está relacionada com a massa do corpo e com a força aplicada.

As teorias clássicas de espaço, tempo e matéria foram trazidas ao auge nas equações de James Clerk Maxwell para o eletromagnetismo.[1] Este elegante sistema de equações unificou duas das forças da natureza, anteriormente julgadas distintas. Em lugar de eletricidade e magnetismo, havia um único campo, o eletromagnético. Um campo permeia a totalidade do espaço, como se o universo estivesse preenchido por algum tipo de fluido invisível. Em cada ponto do espaço podemos medir a intensidade, direção e sentido do campo, como se o fluido estivesse escoando em padrões matemáticos. Para alguns propósitos, o campo eletromagnético

pode ser dividido em dois componentes, o elétrico e o magnético. Mas um campo magnético em movimento gera um campo elétrico, e vice-versa, então, quando se trata de dinâmica, os dois campos devem ser combinados em um único campo mais complexo.

Essa imagem confortável do mundo físico, no qual os conceitos científicos fundamentais mostram uma semelhança próxima às coisas que os nossos sentidos percebem, mudou drasticamente nos primeiros anos do século XX. Naquele momento, os físicos começaram a perceber que em escalas muito pequenas, pequenas demais para serem observadas em qualquer microscópio disponível na época, a matéria é muito diferente do que todo mundo tinha imaginado. Físicos e químicos começaram a levar a sério uma teoria bastante esquisita com mais de dois milênios de idade, remontando às elucubrações filosóficas de Demócrito na Grécia antiga e outros eruditos na Índia. Era a ideia de que, embora o mundo parecesse ser formado de incontáveis materiais diferentes, toda a matéria é composta de minúsculas partículas: os átomos. A palavra vem do grego e significa "indivisível".

Os químicos do século XIX acharam evidência indireta para os átomos: os elementos que se combinam para formar moléculas mais complexas o fazem em proporções bem específicas, muitas vezes próximas a números inteiros. John Dalton formulou essas observações como sua lei das proporções múltiplas, e apresentou os átomos como explicação. Se cada composto químico consistia em números fixos de átomos de vários tipos, esse tipo de relação apareceria automaticamente. Por exemplo, sabemos agora que cada molécula de dióxido de carbono consiste em dois átomos de oxigênio e um átomo de carbono, então os números de átomos estarão na razão dois para um. Contudo, há complicações: átomos diferentes têm massas diferentes, e muitos elementos ocorrem como moléculas formadas por vários átomos – por exemplo, a molécula de oxigênio é composta por dois átomos de oxigênio. Se você não entende o que se passa, pensará que o átomo de oxigênio tem o dobro da massa que efetivamente tem. E alguns aparentes elementos são na verdade misturas de diferentes "isótopos" – estruturas atômicas. Por exemplo, o cloro ocorre na natureza como uma mistura de duas formas estáveis, agora chamadas de cloro-35 e cloro-37, em proporções de cerca de 76% e 24%, respectivamente. Assim, o "peso atômico" observado do cloro é 35,45, o que, nos estágios incipientes da teoria atômica, era interpretado como "o átomo de cloro é composto de 35,5 átomos de hidrogênio". E isso significa que um átomo não é indivisível.

Com a aurora do século XX, a maioria dos cientistas ainda sentia que o salto para a teoria atômica era grande demais e a evidência numérica, fraca demais para justificar sua adoção.

Alguns cientistas, sobretudo Maxwell e Ludwig Boltzmann, estavam à frente da curva, convencidos de que gases são coleções de moléculas esparsamente distribuídas e que as moléculas são feitas juntando-se átomos. O que convenceu a maioria de seus colegas parece ter sido a explicação de Albert Einstein para o movimento browniano, oscilações erráticas de diminutas partículas suspensas num líquido, que eram visíveis ao microscópio. Einstein concluiu que essas oscilações deviam ser causadas por colisões com moléculas do líquido movendo-se aleatoriamente, e realizou alguns cálculos quantitativos para apoiar esta visão. Jean Perrin confirmou essas previsões experimentalmente em 1908. A possibilidade de ver o efeito das alegadas partículas invisíveis de matéria, e fazer previsões quantitativas, carregava mais convicção do que devaneios filosóficos e curiosidade numerológica. Em 1911, Amedeo Avogadro resolveu o problema com isótopos, e a existência de átomos transformou-se em consenso científico.

ENQUANTO ISSO ACONTECIA, alguns cientistas começaram a dar-se conta de que os átomos não são indivisíveis. Eles têm algum tipo de estrutura, e é possível arrancar pedacinhos deles. Em 1897, Joseph John Thomson estava fazendo experimentos com os chamados raios catódicos, e descobriu que podia se fazer com que os átomos emitissem partículas ainda menores, os elétrons. E não só isso: os átomos de diferentes elementos emitiam as mesmas partículas. Aplicando um campo magnético, Thomson mostrou que os elétrons transportam uma carga elétrica negativa. Como um átomo é eletricamente neutro, devia haver também alguma parte do átomo com carga positiva, levando Thomson a propor o modelo do pudim de passas: um átomo é como um pudim carregado positivamente recheado de passas negativamente carregadas. Mas em 1909, um dos ex-alunos de Thomson, Ernest Rutherford, realizou experimentos mostrando que a maior parte da massa de um átomo está concentrada perto de seu centro. Os pudins não são assim.

Como podem os experimentos sondar regiões tão minúsculas do espaço? Imagine um pedaço de terra, que pode ou não ter construções ou outras estruturas. Você não tem autorização para entrar na área, e está

escuro como breu e não é possível enxergar o que há ali. Entretanto, você tem um rifle e muitas caixas de munição. Pode atirar balas ao acaso em direção ao terreno e observar a direção por onde elas saem. Se o terreno for como um pudim de passas, a maioria das balas vai atravessar direto. Se ocasionalmente você tiver de se abaixar quando uma bala ricocheteia de volta na sua direção, há alguma coisa bastante sólida por ali. Observando a frequência com que a bala sai num certo ângulo, você pode avaliar o tamanho do objeto sólido.

As balas de Rutherford eram partículas alfa, núcleos de átomos de hélio, e seu pedaço de terra era uma delgada lâmina de ouro. O trabalho de Thomson mostrara que as passas de elétrons têm massa muito pequena, então quase toda a massa do átomo devia ser encontrada no pudim. Se o pudim não estivesse encaroçado, a maioria das partículas alfa deveria atravessá-lo direto, com muito poucas sendo desviadas, e em um ângulo grande. Em vez disso, uma proporção pequena, mas significativa, sofreu desvios grandes. Então a imagem do pudim de passas não servia. Rutherford sugeriu uma metáfora diferente, que nós ainda usamos informalmente apesar de ter sido superada por imagens mais modernas: o modelo planetário. Um átomo é como o sistema solar: tem um enorme núcleo central, seu sol, ao redor do qual orbitam elétrons como planetas. Logo, assim como o sistema solar, o interior de um átomo é na maior parte espaço vazio.

Rutherford foi além, encontrando evidência de que o núcleo é composto por dois tipos distintos de partículas: prótons, com carga positiva, e nêutrons, com carga nula. Os dois têm massas muito semelhantes e são cerca de 1.800 vezes mais pesados do que um elétron. Os átomos, longe de serem indivisíveis, são feitos de partículas subatômicas ainda menores. Essa teoria explica a numerologia inteira dos elementos químicos: o que está sendo contado são os números de prótons e nêutrons. E também explica os isótopos: somando ou subtraindo alguns nêutrons, a massa muda, mas a carga permanece zero e deixa o número de elétrons – igual ao número de prótons – intacto. As propriedades químicas de um átomo são basicamente controladas por seus elétrons. Por exemplo, o cloro-35 possui dezessete prótons, dezessete elétrons e dezoito nêutrons; o cloro-37 tem dezessete prótons, dezessete elétrons e vinte nêutrons. A cifra 35,45 surge porque o cloro natural é uma mistura desses dois isótopos.

No início do século XX havia surgido uma nova teoria, aplicável à matéria em escalas de partículas subatômicas. Era a mecânica quântica, e,

uma vez tendo se tornado acessível, a física jamais voltaria a ser a mesma. A mecânica quântica previu uma série de novos fenômenos, muitos dos quais foram rapidamente observados em laboratório. Explicou uma porção de observações estranhas e anteriormente desconcertantes. Predisse a existência de novas partículas fundamentais. E nos disse que a imagem clássica do universo em que vivemos, apesar de antes estar em excelente acordo com as observações, está errada. Nossas percepções em escala humana são modelos pobres da realidade em seu nível mais fundamental.

Em física clássica, a matéria é feita de partículas e a luz é uma onda. Em mecânica quântica, a luz também é uma partícula, o fóton; e, ao contrário, por exemplo, os elétrons podem às vezes se comportar como ondas. A linha divisória antes nítida e precisa entre ondas e partículas não só ficou borrada, como sumiu totalmente, substituída pela dualidade onda/partícula. O modelo planetário do átomo não funcionava muito bem se tomado ao pé da letra, de modo que uma nova imagem surgiu. Em vez de orbitar em torno do núcleo como planetas, os elétrons formam uma nuvem difusa centrada no núcleo, uma nuvem não feita de matéria, mas de probabilidades. A densidade da nuvem corresponde à probabilidade de encontrar um elétron naquele local.

Assim como prótons, nêutrons e elétrons, os físicos conheciam mais uma partícula subatômica, o fóton. Logo surgiram outras. Uma aparente falha na lei da conservação da energia levou Wolfgang Pauli a propor um modo de remendá-la postulando a existência do neutrino, uma nova partícula invisível e praticamente indetectável que forneceria a energia que faltava. Era detectável apenas o bastante para sua existência ser confirmada em 1956. Isso abriu as comportas: em pouco tempo havia píons, múons e káons, os últimos descobertos observando-se os raios cósmicos. Nascia a física de partículas, dando continuidade ao método de Rutherford de sondar as escalas espaciais incrivelmente pequenas envolvidas: para descobrir o que há no interior de algo, jogue um monte de coisas e observe o que rebate e volta. Aceleradores de partículas cada vez maiores – na verdade, as armas que disparavam as balas – foram construídos e operados. O acelerador linear de Stanford tinha três quilômetros de comprimento. Para evitar ter de construir aceleradores cujos comprimentos ligassem continentes, foram transformados em círculos, com as partículas podendo dar voltas e mais voltas um enorme número de vezes com velocidades colossais. Isso complicou a tecnologia, porque partículas movendo-se em círculos irradiam energia, mas havia

meios de reparar.

O primeiro fruto desses trabalhos foi um crescente catálogo de partículas alegadamente fundamentais. Enrico Fermi expressou sua frustração: “Se eu conseguisse lembrar os nomes de todas essas partículas, seria um botânico.” Mas de vez em quando novas ideias da teoria quântica voltavam a mexer na lista, à medida que eram propostos novos tipos de partículas cada vez menores para unificar as estruturas já observadas.

A MECÂNICA QUÂNTICA em seus primórdios aplicava-se a coisas individuais como partículas ou ondas. Ninguém conseguia descrever por meio dela uma boa analogia de um campo. Era impossível ignorar essa lacuna porque as partículas (descritíveis pela mecânica quântica) podiam interagir, e de fato interagiam, com campos (não descritíveis pela mecânica quântica). Era como querer descobrir como os planetas do sistema solar se movem conhecendo as leis de Newton do movimento (como as massas se movem quando são aplicadas forças), mas sem conhecer sua lei da gravidade (o que são essas forças).

Havia outro motivo para se querer modelar campos, em vez de simplesmente partículas. Graças à dualidade onda/partícula, eles estão intimamente relacionados. Uma partícula é essencialmente uma fatia de campo comprimida. Um campo é um mar de partículas fortemente apertadas umas contra as outras. Os dois conceitos são inseparáveis. Infelizmente, os métodos desenvolvidos até aquela data baseavam-se em partículas como diminutos pontos, e não se estendiam de forma coerente para os campos. Não era possível simplesmente grudar um monte de partículas umas nas outras e chamar o resultado de campo, porque as partículas *interagem* entre si.

Imagine uma multidão de pessoas… bem, um campo. Talvez estejam em um concerto de rock. Vista por um helicóptero que passa, a multidão parece um líquido, esparramando-se pelo campo – muitas vezes literalmente, por exemplo no Festival de Glastonbury, famoso por ter se tornado um mar de lama. Lá embaixo, no chão, fica claro que o líquido é na verdade uma massa fervilhante de partículas individuais: pessoas. Ou talvez densos aglomerados de pessoas, como quando amigos andam juntos, formando uma unidade indivisível, ou quando grupos de estranhos se juntam para um propósito comum, como, por exemplo, ir ao bar. Mas não se consegue modelar a multidão acuradamente somando o que as pessoas

fariam se estivessem sozinhas. Quando um grupo se dirige ao bar, bloqueia o caminho para outro grupo. Os dois grupos se chocam e se empurram. Estabelecer uma teoria quântica de campo efetiva é como fazê-lo quando as pessoas são funções de onda quânticas localizadas.

No fim da década de 1920, esse tipo de raciocínio havia convencido os físicos de que, por mais dura que a tarefa pudesse ser, a mecânica quântica deveria ser estendida de maneira a tratar de campos além de partículas. O lugar natural para começar era o campo eletromagnético. De algum modo os componentes elétrico e magnético desse campo precisavam ser quantizados: reescritos no formalismo da mecânica quântica. Matematicamente, esse formalismo não era familiar e não muito físico. Coisas observáveis – que podiam ser medidas – não eram mais representadas usando-se os bons e velhos números. Em vez disso, correspondiam a operadores num espaço de Hilbert: regras matemáticas para manipular ondas. Esses operadores violavam as premissas usuais da mecânica clássica. Se você multiplicar dois números entre si, o resultado será o mesmo, não importando quem vier primeiro. Por exemplo, 2 × 3 e 3 × 2 são a mesma coisa. Essa propriedade, chamada comutativa, falha para muitos pares de operadores, assim como calçar as meias e depois os sapatos não tem o mesmo efeito que calçar os sapatos e depois as meias. Números são criaturas passivas, operadores são ativos. A primeira ação executada prepara o terreno para a outra.

A comutatividade é uma propriedade matemática muito agradável. Sua ausência traz um pouco de aborrecimento, e este é exatamente um dos motivos para a quantização dos campos tornar-se algo tão traiçoeiro. Não obstante, às vezes pode ser feita. O campo eletromagnético foi quantizado em uma série de etapas, a começar pela teoria do elétron de Dirac, em 1928, e completada por Sin-Itiro Tomonaga, Julian Schwinger, Richard Feynman e Freeman Dyson entre o final e o início dos anos 1940 e 1950, respectivamente. A teoria resultante é conhecida como eletrodinâmica quântica.

O ponto de vista que foi usado ali sugeriu um método que poderia funcionar de maneira mais genérica. A ideia subjacente voltava diretamente a Newton. Quando os matemáticos tentaram resolver as equações fornecidas pela lei de Newton, descobriram alguns artifícios gerais bastante úteis, conhecidos como leis da conservação. Quando um sistema de massas se move, algumas grandezas permanecem constantes. A mais familiar é a energia, que vem em dois sabores: cinética e potencial. A

energia cinética está relacionada com a velocidade com que o corpo se move, e a energia potencial é o trabalho realizado pelas forças. Quando uma rocha é empurrada da beira de um penhasco, ela troca energia potencial, devido à gravidade, por energia cinética; em linguagem comum, ela cai e acelera. Outras grandezas conservativas são a quantidade de movimento, que é massa vezes velocidade, e o momento angular, que está relacionado com a taxa de giro do corpo. Essas grandezas conservativas relacionam as diversas variáveis usadas para descrever o sistema, e portanto reduzem sua quantidade. Isso ajuda a resolver as equações, como vimos para o problema dos dois corpos no Capítulo 8.

Por volta da primeira década do século XX a fonte dessas leis de conservação já fora compreendida. Emmy Noether provou que toda grandeza conservativa corresponde a um grupo contínuo de simetrias das equações. Uma simetria é uma transformação matemática que deixa as equações inalteradas, e todas as simetrias formam um grupo, sendo a operação "faça uma transformação, depois a outra". Um grupo contínuo é um grupo de simetrias definidas por um único número real. Por exemplo, a rotação em torno de um eixo é uma simetria, e o ângulo de rotação pode ser qualquer número real, então as rotações – por meio de qualquer ângulo – em torno de um eixo dado formam uma família contínua. Aqui a grandeza conservativa associada é o momento angular. Da mesma forma, a quantidade de movimento é a grandeza conservativa associada com a família de translações em determinada direção. E a energia? É a grandeza conservativa correspondente a simetrias de tempo – as equações são as mesmas para todos os instantes de tempo.

QUANDO OS FÍSICOS TENTARAM unificar as forças básicas da natureza, ficaram convencidos de que as simetrias eram a chave. A primeira dessas unificações foi a de Maxwell, combinando eletricidade e magnetismo num único campo eletromagnético. Maxwell obteve essa unificação sem considerar simetria, mas logo ficou claro que suas equações possuem um tipo notável de simetria que não fora anteriormente notado: simetria de calibre, ou simetria gauge. E esta parecia uma alavanca estratégica capaz de abrir teorias quânticas de campo mais gerais.

Rotações e translações são simetrias globais: aplicam-se uniformemente através de todo o espaço e tempo. Uma rotação em torno de um eixo gira cada ponto do espaço de um mesmo ângulo. Simetrias de

calibre são diferentes: são simetrias locais, que podem variar de ponto para ponto no espaço. No caso do eletromagnetismo, essas simetrias locais são mudanças de fase. Uma oscilação local do campo eletromagnético tem tanto amplitude (seu tamanho) como fase (o tempo que leva para atingir seu pico). Se pegarmos uma solução das equações de campo de Maxwell e mudarmos a fase em cada ponto, obtemos outra solução, contanto que façamos a alteração compensatória para a descrição do campo, incorporando uma carga eletromagnética local.

Simetrias de calibre foram introduzidas por Hermann Weyl numa tentativa abortada de conseguir uma nova unificação, do eletromagnetismo com a relatividade geral. Ou seja, das forças eletromagnética e gravitacional. O nome foi usado devido a um mal-entendido: ele imaginou que as simetrias locais corretas deviam ser mudanças na escala espacial, ou no calibre, "gauge". Essa ideia não deu certo, mas o formalismo da mecânica quântica levou Vladimir Fock e Fritz London a introduzir um tipo diferente de simetria local. A mecânica quântica é formulada usando números complexos, não só números reais, e toda função de onda quântica tem uma fase complexa. As simetrias locais relevantes giram de fase através de qualquer ângulo no plano complexo. De maneira abstrata, esse grupo de simetrias consiste em todas as rotações, mas em coordenadas complexas estas são "transformações unitárias" (U) em um espaço com uma dimensão complexa (1), então o grupo formado por essas simetrias é designado por U(1). Aqui o formalismo não é apenas um jogo matemático abstrato: ele permitiu aos físicos escrever, e depois resolver, as equações para partículas quânticas carregadas movendo-se num campo eletromagnético. Nas mãos de Tomonaga, Schwinger, Feynman e Dyson esse ponto de vista levou à primeira teoria quântica de campo relativista do eletromagnetismo: a eletrodinâmica quântica. A simetria sob o grupo de calibre U(1) foi fundamental para seu trabalho.

O passo seguinte, unificar a eletrodinâmica quântica com a força nuclear fraca, foi conseguido por Abdus Salam, Sheldon Glashow, Steven Weinberg, entre outros, nos anos 1960. Juntamente com o campo eletromagnético com sua simetria de calibre U(1), introduziram campos associados com quatro partículas fundamentais, os chamados bósons W^+, W^0, W^- e B^0. As simetrias de calibre deste campo, que efetivamente giravam combinações dessas partículas para produzir outras combinações, formam outro grupo, chamado SU(2) – transformações unitárias (U) num espaço complexo bidimensional (2) que também são especiais (S), uma simples

condição técnica. O grupo de calibre combinado é portanto U(1) × SU(2), onde o "×" indica que os dois grupos atuam independentemente nos dois campos. O resultado, chamado teoria eletrofraca, requeria uma difícil inovação matemática. O grupo U(1) para a eletrodinâmica quântica é comutativo: aplicando-se duas transformações de simetria em sequência, obtém-se o mesmo resultado, não importando que transformação foi aplicada primeiro. Essa agradável propriedade torna a matemática muito mais simples, mas falha para SU(2). Essa foi a primeira aplicação de uma teoria de calibre não comutativa.

A forte força nuclear entra em jogo quando consideramos a estrutura interna de partículas como prótons e nêutrons. O grande avanço nessa área foi motivado por um curioso padrão matemático numa classe particular de partículas, chamadas hádrons. O padrão era conhecido como caminho dos oito preceitos e inspirou a teoria da cromodinâmica quântica, que postulou a existência de partículas ocultas chamadas quarks, usando-as como componentes básicos para o grande zoológico de hádrons.

No Modelo Padrão, tudo no universo é feito a partir de dezesseis partículas genuinamente fundamentais, cuja existência foi confirmada por experimentos em aceleradores. Mais uma décima sétima, pela qual o Grande Colisor de Hádrons está atualmente procurando. Das partículas conhecidas por Rutherford, apenas duas continuam sendo fundamentais: o elétron e o fóton. O próton e o nêutron, por outro lado, são feitos de quarks. O nome foi cunhado por Murray Gell-Mann, que pretendeu uma rima com "cork" (rolha). Ele se deparou com uma passagem no *Finnegans Wake*, de James Joyce:

Three quarks for Muster Mark!
Sure he has not got much of a bark
And sure any he has it's all beside the mark.

Seria de esperar uma pronúncia rimando com "mark", mas Gell-Mann achou um jeito de justificar sua intenção. Atualmente ambas as pronúncias são comuns.

O Modelo Padrão visualizou seis quarks, divididos em pares. Eles têm nomes curiosos: up/down, charmed/strange e top/bottom.[c] Há seis léptons, também em pares: elétron, múon e táuon (hoje, em geral, chamado simplesmente táu). E seus correspondentes neutrinos. Essas doze

partículas são mantidas unidas por meio de forças, que são de quatro tipos: gravidade, eletromagnetismo, força nuclear forte e força nuclear fraca. Deixando de fora a gravidade, que ainda não foi plenamente conciliada com o quadro quântico, isso nos dá três forças. Em física de partículas, as forças são produzidas por troca de partículas, que “transportam” ou “mediam” a força. A analogia usual é a de dois jogadores de tênis mantidos juntos pela sua mútua atenção à bola. O fóton media a força eletromagnética, os bósons Z e W mediam a força nuclear fraca, e o glúon media a força nuclear forte. Bem, tecnicamente ele media a força da cor, que mantém os quarks unidos, e a força forte é o que observamos como resultado. O próton consiste em dois quarks up mais um quark down; o nêutron consiste em um quark up mais dois quarks down. Em cada uma dessas partículas, os quarks são mantidos juntos por glúons. Os quatro portadores de força são conhecidos coletivamente como bósons, em honra a Chandra Bose. A distinção entre férmions e bósons é importante: eles têm diferentes propriedades estatísticas. A Figura 44 (esquerda) mostra o catálogo resultante de partículas conjecturalmente fundamentais. A Figura 44 (direita) mostra como formar um próton e um nêutron a partir de quarks.

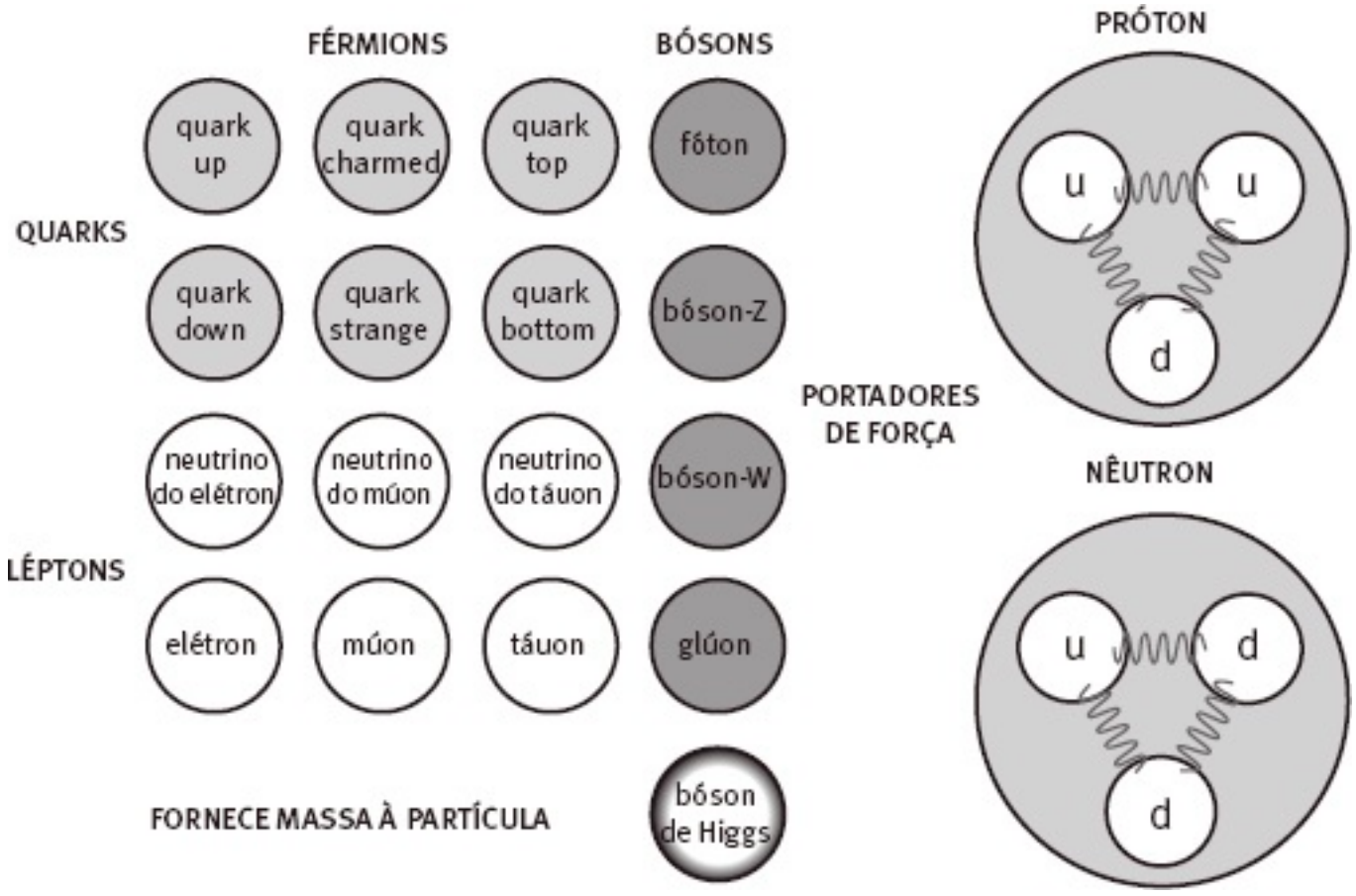


FIGURA 44 *Esquerda*: As dezessete partículas do Modelo Padrão. *Direita*: Como formar um próton e um nêutron a partir de quarks. *Direita no alto*: Próton = dois quarks up + um quark down. *Direita*

embaixo: Nêutron = um quark up + dois quarks down.

O bóson de Higgs completa este quadro explicando por que as outras dezesseis partículas do Modelo Padrão possuem massas não nulas. Ele foi batizado em honra a Peter Higgs, um dos físicos que sugeriram essa ideia. Outros envolvidos são Philip Anderson, François Englert, Robert Brout, Gerald Guralnik, Carl Hagen e Thomas Kibble. O bóson de Higgs é a partícula encarnação de um hipotético campo quântico, o campo de Higgs, com uma característica incomum, porém vital: no vácuo, o campo é diferente de zero. As outras dezesseis partículas são influenciadas pelo campo de Higgs, que faz com que se comportem como se tivessem massa.

Em 1993, David Miller, respondendo a um desafio do ministro da Ciência britânico William Waldegrave, apresentou uma surpreendente analogia: um coquetel. As pessoas espalham-se de maneira uniforme pela sala quando o convidado de honra (uma ex-primeira-ministra) entra. Imediatamente todo mundo se reúne ao seu redor. À medida que ela se movimenta pela sala, diferentes pessoas entram e saem do ajuntamento, e o grupo móvel lhe confere massa adicional, fazendo com que lhe seja mais difícil parar. Este é o mecanismo de Higgs. Agora imagine um boato passando pela sala, com as pessoas se aglomerando para ouvir a notícia. Esse aglomerado é o bóson de Higgs. Miller acrescentou: “Poderia haver um mecanismo de Higgs e um campo de Higgs pelo nosso universo sem haver um bóson de Higgs. A próxima geração de colisores vai resolver isso.” Ele parece ter identificado o bóson de Higgs, mas o campo de Higgs necessita de mais trabalho.

A cromodinâmica quântica é outra teoria de calibre, dessa vez com o grupo de calibre SU(3). Como a notação sugere, as transformações agora atuam em um espaço complexo tridimensional. A unificação do eletromagnetismo, força fraca e força forte foi consequência. Ela assume a existência de três campos quânticos, um para cada força, com grupos de calibre U(1), SU(2) e SU(3), respectivamente. Combinando os três obtém-se o Modelo Padrão, com grupo de calibre $U(1) \times SU(2) \times SU(3)$. Estritamente falando, as simetrias SU(2) e SU(3) são aproximadas; acredita-se que se tornam exatas em energias muito altas. Assim, seu efeito sobre as partículas que compõem o nosso mundo correspondem a simetrias “quebradas” – vestígios da estrutura que permanecem quando o sistema ideal perfeitamente simétrico é sujeito a pequenas perturbações.

Todos os três grupos contêm famílias contínuas de simetrias: uma

família dessas de U(1), três para SU(2) e oito para SU(3). Associadas a elas estão várias grandezas conservativas. As simetrias da mecânica newtoniana mais uma vez fornecem energia, momento linear e momento angular. As grandezas conservativas para as simetrias de calibre U(1) × SU(2) × SU(3) são vários "números quânticos", que caracterizam partículas. Esses são análogos a essas grandezas como spin e carga, mas aplicam-se aos quarks; possuem nomes como carga de cor, isospin e hipercarga. Finalmente, há algumas grandezas conservativas adicionais para U(1): números quânticos para os seis léptons, tais como número de elétron, número de múon e número de táuon. O desfecho é que as simetrias das equações do Modelo Padrão, via teorema de Noether, explicam todas as variáveis físicas essenciais das partículas fundamentais.

A MENSAGEM IMPORTANTE para a nossa história é a estratégia global e o resultado. Para unificar teorias físicas, encontre suas simetrias e unifiqueas. Então monte uma teoria adequada com esse grupo combinado de simetrias. Não estou sugerindo que o processo seja simples; na realidade ele é bastante complexo. Mas até aqui foi assim que a teoria quântica de campo foi desenvolvida, e apenas uma das nossas quatro forças da natureza atualmente cai fora do seu escopo: a gravidade.

Não só o teorema de Noether explica as principais variáveis físicas associadas com as partículas fundamentais: foi assim que muitas das simetrias subjacentes foram encontradas. Os físicos trabalharam de trás para a frente a partir de números quânticos inferidos e observados para deduzir quais simetrias o modelo deveria ter. Então escreveram equações adequadas com essas simetrias, e confirmaram que essas equações ajustavam-se muito de perto à realidade. No momento, esse passo final requer a escolha dos valores de dezenove parâmetros – números que precisam ser conectados às equações para fornecer resultados quantitativos. Nove desses números são massas de partículas específicas: todos os seis quarks e o elétron, o múon e o táu. O resto são coisas mais técnicas, como misturar ângulos e acoplamentos de fase. Dezessete desses parâmetros são conhecidos a partir de experimentos, mas dois não são; eles descrevem o ainda hipotético *campo* de Higgs. Mas agora há uma boa perspectiva de medi-los, porque os físicos sabem onde procurar.

As equações empregadas nessas teorias pertencem a uma classe geral de teorias de campo de calibre, conhecidas como teorias Yang-Mills. Em

1954, Chen-Ning Yang e Robert Mills tentaram desenvolver teorias de calibre para explicar a força forte e as partículas a ela associadas. Suas primeiras tentativas enfrentaram dificuldades quando o campo foi quantizado, pois isso exigiu que as partículas tivessem massa zero. Em 1960, Jeffrey Goldstone, Yoichiro Nambu e Giovanni Jona-Lasinio descobriram um meio de contornar o problema: começar com uma teoria que predissesse partículas sem massa, e então modificá-la quebrando algumas das simetrias. Ou seja, alterar um pouco as equações introduzindo novos termos assimétricos. Quando essa ideia foi usada para modificar a teoria de Yang-Mills, as equações resultantes tiveram um bom desempenho tanto na teoria eletrofraca como na cromodinâmica quântica.

Yang e Mills assumiram que o grupo de calibre era um grupo unitário especial. Para aplicações em partícula era ou SU(2) ou SU(3), o grupo unitário especial para duas ou três dimensões complexas, mas o formalismo funcionava para qualquer número de dimensões. Sua teoria ataca de frente uma dificuldade matemática complexa, mas inevitável. O campo eletromagnético é, sob um aspecto, enganosamente simples: suas simetrias de calibre comutam. Ao contrário da maioria dos operadores quânticos, a ordem em que as fases são alteradas não afeta as equações. Os físicos estavam de olho em uma teoria de campo quântica para partículas subatômicas. Ali, o grupo de calibre era não comutativo, o que tornava as equações quantizadoras muito difíceis.

Yang e Mills tiveram êxito usando uma representação diagramática de interação de partículas introduzida por Richard Feynman. Qualquer estado quântico pode ser pensado como uma superposição de inúmeras interações de partículas. Por exemplo, mesmo o vácuo envolve pares de partículas e antipartículas piscando para dentro da existência e então desaparecendo novamente numa piscada. Uma simples colisão entre duas partículas divide-se em uma espantosa dança de aparições temporárias e desaparições de partículas intermediárias, oscilando para a frente e para trás, repartindo-se e combinando-se. O que salva o dia é uma combinação de duas coisas. As equações de campo podem ser quantizadas para cada diagrama de Feynman específico, e todas essas contribuições podem ser somadas para representar o efeito de toda a interação. Além disso, os diagramas mais complicados são raros, então não contribuem muito para a soma. Mesmo assim, há um problema sério. A soma, interpretada diretamente, é infinita. Yang e Mills acharam uma maneira de "renormalizar" o cálculo de modo que uma infinidade de termos que

realmente não importam fosse removida. O que restava era uma soma finita, e seu valor combinava muito de perto com a realidade. Essa técnica era totalmente misteriosa quando foi inventada, mas agora faz sentido.

Na década de 1970, os matemáticos entraram em ação e Michael Atiyah generalizou a teoria de Yang-Mills para uma ampla classe de grupos de calibre. Matemática e física passaram a alimentar-se mutuamente, e o trabalho de Edward Witten e Nathan Seiberg sobre teorias topológicas de campo quântico levou ao conceito de supersimetria, no qual todas as partículas conhecidas têm novas contrapartes "supersimétricas": elétrons e selétrons, quarks e squarks. Isso simplificou a matemática e levou a predições físicas. No entanto, essas novas partículas ainda não foram observadas, e algumas provavelmente já deveriam ter se manifestado a essa altura nos experimentos realizados usando o Grande Colisor de Hádrons. O valor matemático dessas ideias está bem-estabelecido, mas sua relevância direta para a física, não. Contudo, elas lançaram uma proveitosa luz sobre a teoria de Yang-Mills.

A teoria quântica de campo é uma das fronteiras da física matemática que se movem mais depressa, sendo assim o Instituto Clay quis incluir algo a respeito da matéria como um dos prêmios do milênio. A hipótese da ausência de massa está firmemente assentada nessa rica área, e aborda um importante tópico matemático ligado à física de partículas. A aplicação dos campos de Yang-Mills para descrever partículas fundamentais em termos de força nuclear forte depende de uma característica quântica específica conhecida como *mass gap*.[d] Em relatividade, uma partícula que viaja na velocidade da luz adquire massa infinita, a menos que sua massa seja zero. O *mass gap* permite que partículas quânticas tenham massas finitas não nulas, mesmo que as ondas clássicas a elas associadas viajem na velocidade da luz. Quando existe um *mass gap*, qualquer estado que não seja o vácuo tem uma energia que excede à do vácuo por algum valor fixo mínimo. Isto é, existe um limite inferior não nulo para a massa de uma partícula.

Experimentos confirmam a existência do *mass gap*, e simulações de computador das equações apoiam a hipótese do *mass gap.* Todavia, não podemos assumir que o modelo se encaixe na realidade e então usar a realidade para verificar características matemáticas do modelo, porque a lógica torna-se circular. Logo, necessita-se de uma compreensão teórica. Um passo fundamental seria uma prova rigorosa de que as versões quânticas da teoria de Yang-Mills existem. A versão clássica (não quântica) é bastante bem entendida nos dias de hoje, mas a análoga quântica é

atormentada pelo problema da renormalização – aquelas incômodas infinidades que precisaram ser espantadas por subterfúgios matemáticos.

Uma abordagem atraente começa transformando o espaço contínuo em um reticulado discreto e escrevendo uma análoga reticulada da equação de Yang-Mills. O ponto principal é então mostrar que à medida que o reticulado se torna cada vez mais fino, aproximando-se de um continuum, essa análoga converge para um objeto matemático bem-definido. Algumas características necessárias da matemática podem ser inferidas a partir da intuição física, e seria possível provar que existe uma teoria quântica de Yang-Mills adequada se essas características puderem ser estabelecidas rigorosamente. A hipótese do *mass gap* envolve uma compreensão mais detalhada de como as teorias de reticulado se aproximam desta teoria hipotética de Yang-Mills. Logo, a existência da teoria e da hipótese do *mass gap* estão estreitamente interligadas.

E é aí que todo mundo encalha. Em 2004, Michael Douglas escreveu um relatório sobre a situação do problema, dizendo: “Até onde sei, não foi feito nenhum avanço neste problema nos últimos anos. Em particular, enquanto foi feito progresso em teorias de campo dimensional inferiores, não sei de nenhum progresso significativo rumo a uma construção matemática rigorosa da teoria quântica de Yang-Mills.” Esta avaliação ainda parece estar correta.

O progresso, porém, tem sido mais impressionante em alguns problemas correlacionados, o que pode lançar alguma luz proveitosa. Teorias de campo quânticas especiais, conhecidas como modelos sigma bidimensionais, são mais tratáveis, e a hipótese do *mass gap* foi estabelecida para um desses modelos. Teorias de campo quântico supersimétricas, envolvendo superparceiras hipotéticas das partículas fundamentais usuais, possuem características matemáticas agradáveis que efetivamente removem a necessidade de renormalização. Físicos como Edward Witten têm feito progresso no sentido de solucionar questões correlatas no caso supersimétrico. A esperança é que alguns dos métodos que surjam desse trabalho possam sugerir novos meios de atacar o problema original. Mas quaisquer que possam ser as implicações físicas, e por mais que a hipótese do *mass gap* venha a ser eliminada, muitos desses desenvolvimentos já enriqueceram a matemática com novos e importantes conceitos e ferramentas.

[c] Literalmente: acima/abaixo, encantado/estranho, topo/fundo. A tradução é apenas curiosidade, pois os nomes costumam ser utilizados em inglês. (N.T.)
[d] *Mass gap* significa literalmente “lacuna de massa”, “falha de massa”, “diferença de massa”, mas o termo é consagrado e utilizado internacionalmente em inglês. (N.T.)

14. Sonhos diofantinos

A conjectura de Birch–Swinnerton-Dyer

NO CAPÍTULO 7 encontramos a *Arithmetica* de Diofanto, e comentei que seis de seus treze livros sobrevivem como cópias em grego. Por volta do ano 400, quando a antiga civilização grega entrou em declínio, Arábia, China e Índia assumiram a tocha da inovação matemática. Estudiosos árabes traduziram muitas das obras clássicas gregas, e com frequência essas traduções são nossa principal fonte histórica de seu conteúdo. O mundo árabe conhecia a *Arithmetica*, e baseou-se nela. Quatro manuscritos arábicos descobertos em 1968 podem ser traduções dos outros livros "que faltam" da *Arithmetica*.

Em alguma época próximo ao fim do século X, o matemático persa al-Karaji fez uma pergunta que poderia facilmente ter ocorrido a Diofanto. Que inteiros podem ocorrer como diferença comum entre três quadrados racionais que formam uma sequência aritmética? Por exemplo, os quadrados inteiros 1, 25 e 49 têm em comum a diferença 24. Isto é, 1 + 24 = 25 e 25 + 24 = 49. Al-Karaji viveu entre 953 e 1029, então pode ter tido acesso a uma cópia da *Arithmetica*, mas a mais antiga tradução conhecida foi feita por Abu'l-Wafā, em 998. Leonard Dickson, que escreveu uma sinopse em três volumes da história da teoria dos números, sugeriu que o problema pode ter se originado em algum momento antes de 972 em algum manuscrito árabe anônimo.

Em linguagem algébrica o problema passa a ser: para quais inteiros d existe um número racional x tal que $x - d$, x e $x + d$ sejam quadrados perfeitos? E pode ser reformulado numa forma equivalente, embora a equivalência não seja tão óbvia: que números inteiros podem ser a área de um triângulo retângulo com lados racionais? Isto é: se a, b e c são racionais e $a^2 + b^2 = c^2$, quais são os possíveis valores inteiros para $^{ab}/_{2}$? Inteiros que satisfaçam essa condição de equivalência são chamados números congruentes. O termo não tem relação com outros usos da palavra "congruente" em matemática, o que o torna ligeiramente confuso para o leitor moderno. Suas origens são explicadas abaixo.

Alguns números não são congruentes: por exemplo, pode-se provar que

1, 2, 3 e 4 não são congruentes. Outros, tais como 5, 6 e 7, são congruentes. De fato, o triângulo 3-4-5 tem área $3 \times {}^{4}/_{2} = 6$, provando que 6 é congruente. Para provar que 7 é congruente, observe que $({}^{24}/_{5})^2$, $({}^{35}/_{12})^2$ e $({}^{337}/_{60})^2$ têm uma diferença comum de 7. Voltarei ao 5 em um instante. Se prosseguirmos caso a caso dessa maneira, obteremos uma longa lista de números congruentes, mas isso lança pouca luz sobre sua natureza. Nenhuma quantidade de construções caso a caso pode provar que um número inteiro específico *não* é congruente. Durante séculos ninguém sabia se 1 era congruente.

Agora sabemos que o problema vai muito além de qualquer coisa que Diofanto pudesse ter solucionado. Na verdade, essa questão enganadoramente simples ainda não foi respondida de maneira completa. O mais perto que chegamos é uma caracterização de números congruentes, descoberta por Jerrold Tunnell em 1983. A ideia de Tunnell fornece um algoritmo para decidir se um dado inteiro pode ou não ocorrer contando-se suas representações como duas diferentes combinações de quadrados. Com um pouco de engenhosidade esse cálculo é viável para inteiros bastante grandes. A caracterização tem apenas uma séria desvantagem: ela nunca se provou correta. Sua validade depende de se resolver um dos problemas do milênio, a conjectura Birch–Swinnerton-Dyer. Essa conjectura fornece um critério para que uma curva elíptica tenha apenas uma quantidade finita de pontos racionais. Encontramos essas equações diofantinas no Capítulo 6, sobre a conjectura de Mordell, e no Capítulo 7, sobre o último teorema de Fermat. Aqui vemos evidência adicional de seu papel proeminente nas fronteiras da teoria dos números.

O PRIMEIRO TRABALHO europeu referindo-se a essas questões foi escrito por Leonardo de Pisa. Ele é mais conhecido por uma sequência de estranhos números que parece ter inventado, e que surgiram num problema aritmético bem pouco realista sobre reprodução de alguns coelhos. São os números de Fibonacci

0 1 1 2 3 5 8 13 21 34 55 89 …

nos quais cada um, após os dois primeiros, é a soma dos dois anteriores. O pai de Leonardo era um funcionário da alfândega chamado Bonaccio, e o apelido famoso significa “filho de Bonaccio”. Não há evidência de que tenha sido usado durante a vida de Leonardo e acredita-se que foi uma invenção

do matemático francês Guillaume Libri no século XIX.[1] Seja como for, os números de Fibonacci possuem muitas propriedades fascinantes, e são amplamente conhecidos. Aparecem até mesmo no suspense criptoconspiratório de Dan Brown, *O código Da Vinci*.

Leonardo introduziu os números de Fibonacci em um livro-texto sobre aritmética, o *Liber Abbaci* (Livro dos cálculos) de 1202, cujos principais objetivos eram chamar a atenção da Europa para a nova notação aritmética dos árabes, baseada nos dígitos 0-9, e demonstrar sua utilidade. A ideia já alcançara a Europa por meio do texto de al-Khwārizmī, de 825, em sua tradução latina *Algoritmi de Numero Indorum* (Sobre o cálculo com numerais indianos), mas o livro de Leonardo foi o primeiro a ser escrito com a intenção específica de promover a incorporação da notação decimal na Europa. Grande parte do livro dedica-se à aritmética prática, sobretudo câmbio de moedas. Mas Leonardo escreveu outro livro, não tão conhecido, que sob muitos aspectos foi um sucessor europeu da *Arithmetica* de Diofanto: o seu *Liber Quadratorum* (Livro dos quadrados).

Como Diofanto, ele apresenta técnicas gerais usando exemplos especiais. Uma delas surge a partir da questão de al-Karaji. Em 1225, o imperador Frederico II visitou Pisa. Ciente da reputação matemática de Leonardo, aparentava ter decidido que seria divertido testá-la num torneio matemático. Tais competições públicas eram comuns na época. Os competidores apresentavam questões uns aos outros. O time do imperador consistia de João de Palermo e Mestre Teodoro. O time de Leonardo consistia de Leonardo. O time do imperador desafiou Leonardo a achar um quadrado que se mantivesse quadrado com a adição ou subtração de 5. Como de hábito, os números deviam ser racionais. Em outras palavras, queriam uma prova de que 5 é um número congruente, mediante a descoberta de um racional específico x para o qual $x - 5$ e $x + 5$ são quadrados. Isso de forma alguma é trivial – a menor solução sendo

$$x = \frac{1681}{144} = \left(\frac{41}{12}\right)^2$$

e neste caso

$$x - 5 = \frac{961}{144} = \left(\frac{31}{12}\right)^2 \quad e \quad x + 5 = \frac{2401}{144} = \left(\frac{49}{12}\right)^2$$

Leonardo encontrou uma solução, e a incluiu no *Liber Quadratorum*. Ele conseguiu a resposta usando uma fórmula geral relacionada com a fórmula de Euclides/Diofanto para trincas pitagóricas. A partir dela obteve três quadrados inteiros com uma diferença comum de 720, ou seja, 31^2, 41^2 e 49^2. Aí dividiu por $12^2 = 144$ para obter três quadrados com uma diferença comum de $720/144$, que é 5.[2] Em termos de trincas pitagóricas, tomemos o triângulo com lados 9, 40 e 41, de área 180, e dividamos por 36 para obter um triângulo de lados $20/3$, $3/2$ e $41/6$. Então sua área é 5.

É em Leonardo que encontramos a palavra latina *congruum* para um conjunto de três quadrados em progressão aritmética. Mais tarde, Euler empregou a palavra *congruere*, "reunir". Os dez primeiros números congruentes, e as correspondentes trincas pitagóricas mais simples, estão listados na Tabela 3. Não há padrões simples evidentes.

A maior parte do progresso inicial nessa questão foi feito por matemáticos islâmicos, que mostraram que os números 5, 6, 14, 15, 21, 30, 34, 65, 70, 110, 154 e 190 são congruentes, junto com dezoito números maiores. A estes, Leonardo, Angelo Genocchi (1855) e André Gérardin (1915) adicionaram 7, 22, 41, 69, 77, além de 43 outros números menores que 1.000. Leonardo afirmou em 1225 que 1 não é congruente, mas não provou. Em 1569, Fermat apresentou a prova. Em 1915, todos os números congruentes menores que 100 haviam sido determinados, mas o problema ganhou terreno lentamente, e em 1980 o status de muitos números menores que 1.000 permaneceu sem solução. A dificuldade pode ser julgada pela descoberta de L. Bastien de que 101 é congruente. As medidas correspondentes aos lados do triângulo retângulo são

d	TRINCA PITAGÓRICA
5	3/2, 20/3, 41/6
6	3, 4, 5
7	24/5, 35/12, 337/60
13	780/323, 323/30, 106921/9690
14	8/3, 63/6, 65/6
15	15/2, 4, 17/2
20	3, 40/3, 41/3
21	7/2, 12, 25/2
22	33/35, 140/3, 4901/105
23	80155/20748, 41496/3485, 905141617/72306780

TABELA 3 Os dez primeiros números congruentes e as correspondentes trincas pitagóricas.

$$\frac{711024064578955010000}{118171431852779451900}$$

$$\frac{396727280603349500 3922}{118171431852779451900}$$

$$\frac{403048492589952000 3922}{118171431852779451900}$$

Ele descobriu esses números em 1914, à mão. Em 1986, agora com computadores em cena, G. Kramarz havia descoberto todos os números congruentes menores que 2.000.

Em algum ponto notou-se que uma equação diferente, mas correlacionada

$$y^2 = x^3 - d^2x$$

tem soluções x, y nos números inteiros se, e somente se, d for congruente.[3] Esta observação é óbvia num sentido: o lado direito da equação é o produto de x, $x - d$ e $x + d$, e se estes forem todos quadrados, então o produto também será. O inverso também é bastante direto. Essa reformulação do problema o coloca diretamente numa área rica e florescente da teoria dos números. Para qualquer d dado esta equação determina que y^2 é igual a um polinômio cúbico em x, e portanto define uma curva elíptica. Assim, o problema dos números congruentes é um caso especial de uma questão que os teóricos dos números adorariam tremendamente responder: quando uma curva elíptica tem ao menos um ponto racional? Essa questão está longe de ser simples, mesmo para o tipo especial de curva elíptica que acabamos de mencionar. Por exemplo, 157 é um número congruente, mas o triângulo retângulo *mais simples* com essa área tem hipotenusa

$$\frac{224403517704336969924557513090667486316094847 2041}{891233226892885958802553517896716357001648083 0}$$

Antes de prosseguir, tomamos emprestado o truque de Leonardo, aquele que o levou de 720 para 5, e o aplicamos em sua plena generalidade. Se multiplicarmos qualquer número congruente d pelo quadrado n^2 de um inteiro n, também obtemos um número congruente. Basta pegar uma trinca pitagórica racional correspondente a um triângulo de área d e multiplicar

os números por n. A área do triângulo fica multiplicada por n^2. O mesmo é verdade se dividirmos os números por n; agora a área fica dividida por n^2. Esse processo dá um número inteiro apenas quando a área tem um fator quadrado, de modo que ao pesquisar números congruentes basta trabalhar com números que não tenham fator quadrado. Os primeiros números sem fator quadrado são

1 2 3 5 6 7 10 11 13 14 15 17 19

Agora podemos formular o critério de Tunnell. Um número ímpar d sem fator quadrado é congruente se, e somente se, o número de soluções inteiras (positivas ou negativas) x, y, z para a equação

$$2x^2 + y^2 + 8z^2 = d$$

for precisamente o dobro do número de soluções para a equação

$$2x^2 + y^2 + 32z^2 = d$$

Um número par d sem fator quadrado é congruente se, e somente se, o número de soluções inteiras x, y, z para a equação

$$8x^2 + 2y^2 + 16z^2 = d$$

for precisamente o dobro do número de soluções para a equação

$$8x^2 + 2y^2 + 64z^2 = d$$

Estes resultados são mais úteis do que podem parecer à primeira vista. Como todos os coeficientes são positivos, as dimensões de x, y, z não podem exceder certos múltiplos da raiz quadrada de d. Assim, o número de soluções é finito, e podem ser encontradas por uma busca sistemática, com alguns atalhos úteis. Eis os cálculos completos para alguns exemplos com d pequeno:

- Se $d = 1$, então as únicas soluções da primeira equação são $x = 0, y = \pm 1, z = 0$. O mesmo vale para a segunda equação. Então ambas as equações têm duas soluções, e o critério não se aplica.
- Se $d = 2$, então as únicas soluções da primeira equação são $x = \pm 1, y = 0, z = 0$. O mesmo vale para a segunda equação. Então ambas as equações têm duas soluções, e o critério não se aplica.

- Se $d = 3$, então as únicas soluções da primeira equação são $x = \pm 1$, $y = \pm 1$, $z = 0$. O mesmo vale para a segunda equação. Então ambas as equações têm quatro soluções, e o critério não se aplica.
- Se $d = 5$ ou 7, então a primeira equação não tem soluções. O mesmo vale para a segunda equação. Como duas vezes zero é zero, o critério é satisfeito.
- Se $d = 6$, temos que usar o critério para números pares. Mais uma vez ambas as equações não têm soluções, e o critério é satisfeito.

Estes cálculos simples mostram que 1, 2, 3, 4 ($= 2^2 \times 1$) não são congruentes, mas 5, 6 e 7 são. A análise pode ser facilmente estendida, e em 2009 uma equipe de matemáticos aplicou o teste de Tunnell para o primeiro trilhão de números, achando exatamente 3.148.379.694 números congruentes. Os pesquisadores verificaram seus resultados executando os cálculos duas vezes, em computadores diferentes, usando algoritmos diferentes elaborados por dois grupos independentes. Bill Hart e Gonzalo Tornaria usaram o computador Selmer na Universidade de Warwick. Mark Watkins, David Harvey e Robert Bradshaw usaram o computador Sage na Universidade de Washington.

No entanto, há uma lacuna em todos esses cálculos. Tunnell provou que se um número d é congruente, então precisa satisfazer o seu critério. Portanto, se o critério falha, o número não é congruente. Isso implica, por exemplo, que 1, 2, 3 e 4 não são congruentes. Todavia, ele foi incapaz de provar o inverso: se um número satisfaz o seu critério, então deve ser congruente. É disso que precisamos para concluir que 5, 6 e 7 são congruentes. Nesses casos particulares encontramos trincas pitagóricas convenientes, mas isso não adianta no caso geral. Tunnell mostrou que esta recíproca é consequência da conjectura Birch–Swinnerton-Dyer – mas esta permanece sem prova.

COMO VÁRIOS dos problemas do milênio, a conjectura Birch–Swinnerton-Dyer é difícil até mesmo de formular. (Você pensa que pode ganhar 1 milhão de dólares fazendo uma coisa fácil? Eu tenho uma ponte espetacular para lhe vender, preço bem baratinho…) No entanto, ela premia a perseverança, porque ao longo do caminho começamos a apreciar a profundidade, e longas tradições históricas, da teoria dos números. Se você olhar com cuidado o nome da conjectura, dois nomes são separados por travessão,

outros dois por hífen. Não é algo conjecturado por três matemáticos chamados Birch, Swinnerton e Dyer, mas por Brian Birch e Peter Swinnerton-Dyer. Sua formulação geral é técnica, mas trata de um assunto básico em equações diofantinas – equações algébricas para as quais buscamos soluções em números inteiros ou racionais. A questão é simples: quando é que elas têm soluções?

No Capítulo 6, que trata da conjectura de Mordell, e no Capítulo 7, sobre o último teorema de Fermat, encontramos alguns dos inventos mais incríveis de toda a matemática, as curvas elípticas. Mordell fez o que na época foi basicamente dar um tiro no escuro, e conjecturou que a quantidade de soluções racionais de uma equação algébrica em duas variáveis depende da topologia da curva complexa associada. Se o genus é 0 – a curva é topologicamente uma esfera – então as soluções são dadas por uma fórmula. Se o genus é 1 – a curva é topologicamente um toro, que equivale a ser uma curva elíptica – então todas as soluções racionais podem ser construídas a partir de uma lista finita conveniente aplicando-se uma estrutura de grupo natural. Se o genus é 2 ou mais – a curva é topologicamente um toro com g furos, com $g \geq 2$ – então o número de soluções é finito. Como vimos, Faltings provou esse notável teorema em 1983.

A característica mais surpreendente das soluções racionais para equações de curvas elípticas é que essas soluções formam um grupo, graças à construção geométrica da Figura 28 no Capítulo 6. A estrutura resultante é chamada grupo de Mordell-Weil da curva, e os teóricos de números adorariam ser capazes de calculá-la. Isso envolve encontrar um sistema de geradores: soluções racionais a partir das quais todas as outras podem ser deduzidas usando-se repetidamente a operação de grupo. Fracassando, gostaríamos ao menos de calcular algumas das características básicas do grupo, tais como seu tamanho. Aqui ainda há muitos detalhes não compreendidos. Às vezes o grupo é infinito, então leva a uma quantidade infinita de soluções racionais; às vezes não é, e a quantidade de soluções racionais é finita. Seria proveitoso poder saber qual é qual. De fato, o que realmente gostaríamos de saber é a estrutura abstrata do grupo.

A prova de Mordell de que uma lista finita gera todas as soluções nos diz que o grupo deve ser constituído por um grupo finito e outro reticulado. Um grupo reticulado consiste de todas as listas de inteiros de algum comprimento finito. Se o comprimento é três, por exemplo, então o grupo consiste de todas as listas (m_1, m_2, m_3) de inteiros, e listas são somadas da

maneira óbvia:

$$(m_1, m_2, m_3) + (n_1, n_2, n_3) = (m_1 + n_1, m_2 + n_2, m_3 + n_3)$$

O comprimento da lista é chamado ordem do grupo (e geometricamente é a dimensão do reticulado). Se a ordem é 0, o grupo é finito. Se a ordem é diferente de zero, o grupo é infinito. Assim, para decidir quantas soluções existem, não precisamos estruturar totalmente o grupo. Tudo de que necessitamos é sua ordem. E é disso que trata a conjectura Birch–Swinnerton-Dyer.

Na década de 1960, logo após o advento dos computadores, a Universidade de Cambridge teve um dos primeiros, chamado Edsac (*electronic delay storage automatic calculator*), em português, "calculador automático de armazenamento com retardo eletrônico", e mostra quanto seus inventores estavam orgulhosos do seu sistema de memória, que enviava ondas sonoras através de tubos de mercúrio e as redirecionava de volta ao começo. Tinha o tamanho de um caminhão grande, e lembro-me muito bem de como o exibiam em 1963. Seus circuitos baseavam-se em milhares de válvulas – tubos de vácuo. Havia enormes painéis de componentes ocupando as paredes, peças de reposição a serem inseridas quando uma válvula da própria máquina estourava. O que acontecia com bastante frequência.

Peter Swinnerton-Dyer estava interessado no lado diofantino das curvas elípticas, e em particular queria entender quantas soluções haveria se a curva fosse substituída por seu análogo num campo finito com um número p primo de elementos. Ou seja, queria estudar o artifício de Gauss de trabalhar "módulo p". E usou o computador para calcular esses números para inúmeros primos, procurando por padrões interessantes.

E aqui está o que ele começou a desconfiar. No começo, seu supervisor John William Scott ("Ian") Cassels estava extremamente cético, mas à medida que mais e mais dados foram entrando, começou a acreditar que poderia haver algo nessa ideia. O que os experimentos de computador de Swinnerton-Dyer sugeriam era o seguinte: os teóricos dos números têm um método padrão que reinterpreta qualquer equação de inteiros comuns em termos de inteiros em algum módulo – lembre-se da "aritmética do relógio" em módulo 12 no Capítulo 2. Como as regras da álgebra se aplicam a esta versão da aritmética, qualquer solução da equação original torna-se solução da equação "reduzida" naquele módulo. Como os números

envolvidos formam uma lista finita – apenas doze números para a aritmética do relógio, por exemplo –, podem-se achar todas as soluções por tentativa e erro. Em particular, podem-se contar quantas soluções existem, para qualquer módulo dado. Soluções em qualquer módulo também impõem condições sobre as soluções inteiras originais, e às vezes podem até provar que tais soluções existem. Então, é reflexo entre os teóricos dos números reduzir equações usando-se vários módulos, e os primos são uma escolha especialmente útil.

Assim, para descobrir alguma coisa acerca de uma curva elíptica, podem-se considerar todos os primos até um limite específico. Para cada primo, podem-se descobrir quantos pontos se encontram sobre a curva, de módulo aquele mesmo primo. Birch notou que os experimentos de computador de Swinnerton-Dyer produzem um interessante padrão caso se divida o número desses pontos pelo referido primo. Então multiplicam-se entre si todas essas frações, para todos os primos menores ou igual ao primo dado, e colocam-se os resultados contra primos sucessivos num gráfico em papel logarítmico. Os dados parecem estar todos nas proximidades de uma linha reta, cuja inclinação é a ordem da curva elíptica. Isso levou a uma fórmula conjecturada para o número de soluções associadas com qualquer módulo primo.[4]

Entretanto, a fórmula não provém da teoria dos números: ela envolve análise complexa, a queridinha do século XIX, que, por algum milagre, é muito mais elegante que a antiquada análise real. No Capítulo 9, sobre a hipótese de Riemann, vimos como a análise estende seus tentáculos em todas as direções, tendo em particular surpreendentes e poderosas ligações com a teoria dos números. A fórmula de Swinnerton-Dyer levou a uma conjectura mais detalhada sobre um tipo de função complexa que mencionei no Capítulo 9, chamada função-*L* de Dirichlet. Essa função é análoga, para as curvas elípticas, à notória função zeta de Riemann. Os dois matemáticos estavam decididamente levando o barco para águas desconhecidas, porque na época não se sabia com certeza que todas as curvas elípticas *tinham* funções-*L* de Dirichlet. Era um puro palpite respaldado por tênues evidências. Mas à medida que cresceu o conhecimento da área, o palpite foi parecendo cada vez mais inspirado. Não era um salto maluco para o desconhecido: era um golpe de larga visão, magnificamente preciso, de refinada intuição matemática. Em vez de se porem de pé sobre ombros de gigantes, Birch e Swinnerton-Dyer colocaram-se sobre seus próprios ombros – gigantes capazes de pairar no

ar.

Uma ferramenta básica em análise complexa é exprimir uma função usando uma série de potências, como um polinômio, mas contendo infinitos termos, usando potências cada vez maiores da variável, que nessa área é tradicionalmente chamada *s*. Para descobrir o que ocorre com uma função perto de um ponto específico, digamos 1, usam-se potências de (*s* − 1). A conjectura de Birch–Swinnerton-Dyer afirma que se a expansão de uma série de potências perto de 1 de uma função-*L* de Dirichlet tem a aparência

$L(C, s) = c(s - 1)^r$ + termos de ordem mais elevada

onde *c* é uma constante não nula, então a ordem da curva é *r*, e vice-versa. Na linguagem da análise complexa, esta formulação toma a forma "$L(C, s)$ tem um zero de ordem *r* em $s = 1$".

O ponto crucial aqui não é a expressão precisa exigida: é que dada qualquer curva elíptica, existe um cálculo analítico, usando uma função complexa relacionada, que nos diz precisamente quantas soluções racionais independentes temos que descobrir para especificá-las todas.

TALVEZ O MODO MAIS SIMPLES de demonstrar que a conjectura de Birch–Swinnerton-Dyer tem conteúdo genuíno seja observar que a maior ordem conhecida é 28. Isto é, existe uma curva elíptica que tem um conjunto de 28 soluções racionais, a partir da qual todas as soluções racionais podem ser deduzidas. Mais ainda, nenhum conjunto menor de soluções racionais é capaz de fazer isso. Embora saiba-se que existem curvas dessa ordem, não foi encontrado nenhum exemplo explícito. A maior ordem para um exemplo explícito é dezoito. A curva, descoberta por Noam Elkies em 2006, é:

$$y^2 + xy = x^3 - 2617596009270588409631170178770120390355643896951$$
$$+ 5106938147613148648974217710037377208977910325389$$
$$7848326$$

Da forma como aparece, ela não está na forma padrão "y^2 = cúbica em x", mas pode ser transformada nessa forma à custa de tornar os números

ainda maiores. Acredita-se que a ordem pode ser indefinidamente grande, mas isso não foi provado. Pelo que sabemos, a ordem nunca pode exceder algum tamanho fixo.

A maior parte do que podemos provar refere-se a curvas de ordem 0 ou 1. Quando a ordem é 0, existem finitas soluções racionais. Quando é 1, uma solução específica leva a quase todo o resto, talvez com um número finito de exceções. Esses dois casos incluem todas as curvas elípticas da forma $y^2 = x^3 + px$ quando p é um primo da forma $8k + 5$ (tal como 13, 29, 37, e assim por diante). Conjectura-se que nesses casos a ordem é sempre 1, o que implica que existem infinitas soluções racionais. Andrew Bremner e Cassels provaram que isso é verdade para todos os primos até 1.000. Pode ser complicado encontrar soluções que levem a quase todas as outras, mesmo que a ordem seja conhecida, e pequena. Eles descobriram que quando p = 877 a solução mais simples desse tipo é o número racional

$$\frac{375494528127162193105504069942092792346201}{6215987776871505425463220780697238044100}$$

Muitos teoremas relacionados com a conjectura de Birch–Swinnerton-Dyer foram provados, geralmente com premissas muito técnicas, mas o progresso rumo à solução tem sido relativamente rápido. Em 1976, Coates e Wiles descobriram o primeiro indício de que a conjectura podia ser verdadeira. Provaram que um tipo especial de curva elíptica tem ordem 0 se a função-L de Dirichlet não desaparece em 1. Para tal curva elíptica, o número de soluções racionais para a equação diofantina é finito, talvez zero – e pode-se deduzir isso da correspondente função-L. Desde então, tem havido inúmeros avanços técnicos, ainda limitados em sua maior parte às ordens 0 e 1. Em 1990, Victor Kolyvagin provou que a conjectura de Birch–Swinnerton-Dyer é verdadeira para as ordens 0 e 1.

Conjecturas mais detalhadas, com bastante ajuda do computador, relacionam a constante c da conjectura de Birch–Swinnerton-Dyer a vários conceitos da teoria dos números. Há análogas – igualmente enigmáticas – para os campos numéricos algébricos. Sabe-se também, num sentido preciso, que a maioria das curvas elípticas tem ordem 0 ou 1. Em 2010, Manjul Bhargava e Arul Shankar anunciaram ter provado que a ordem média de uma curva elíptica é no máximo $7/6$. Se isso e outros teoremas recentemente anunciados sustentaram-se sob escrutínio preciso, a conjectura de Birch–Swinnerton-Dyer é verdadeira para uma proporção não nula de todas as curvas elípticas. No entanto, estas são as mais simples

e não representam de fato as curvas com estrutura mais rica: ordem 2 ou mais. Estas estão envoltas em mistério quase total.

15. Ciclos complexos

A conjectura de Hodge

ALGUMAS ÁREAS DA MATEMÁTICA podem estar relacionadas, de forma bastante direta, com acontecimentos e preocupações do dia a dia. Não encontramos a equação de Navier-Stokes na nossa cozinha, mas todos sabemos o que são os líquidos e temos ideia de como fluem. Algumas áreas podem estar relacionadas com questões esotéricas na fronteira da ciência: você pode precisar de um doutorado em física matemática para compreender a teoria quântica de campo, mas analogias com eletricidade e magnetismo, ou imagens semissignificativas como "onda de probabilidade", conseguem dar uma boa visão. Algumas podem ser explicadas usando figuras: a conjectura de Poincaré é um bom exemplo. Mas outras desafiam todos esses métodos de tornar acessíveis conceitos abstratos difíceis.

A conjectura de Hodge, formulada pelo geômetra escocês William Hodge em 1950, é uma delas. Não é a prova que causa problemas, porque não existe uma. É a formulação. Aqui está ela tirada do site do Instituto Clay, em forma ligeiramente editada:

Em qualquer variedade algébrica complexa projetiva não singular, qualquer classe de Hodge é uma combinação linear racional de classes de ciclos algébricos.

Está claro que temos trabalho pela frente. As únicas palavras que formam sentido imediato são "em, qualquer, é, uma" e "de". Outras são familiares como as palavras: "variedade, classe, racional, ciclo". Mas as imagens que elas conjuram – escolha no supermercado, uma sala de aula cheia de crianças, pensamento não emotivo, parte de um equipamento dotado de duas rodas e guidão – não são, obviamente, os sentidos que o Instituto Clay tem em mente. O restante é jargão ainda mais evidente. Mas não é jargão pelo simples prazer de ser jargão – nomes complicados para coisas simples. Estes são nomes simples para coisas complicadas. Não há nomes prontos para tais conceitos na linguagem comum, de modo que tomamos alguns emprestados e inventamos outros.

Olhando pelo lado positivo, temos aqui uma oportunidade de verdade (no sentido de: “cara, que bela oportunidade nós temos!”). A conjectura de Hodge é indiscutivelmente mais representativa da matemática real, como é feita pelos matemáticos dos séculos XX e XXI, do que qualquer outro tópico deste livro. Abordando-a da maneira correta, obtemos uma percepção valiosa de quanto está de fato conceitualmente avançada a fronteira da matemática. Comparada com a matemática escolar, é como o monte Everest ao lado de um montículo.

Então é tudo uma viagem mental absurda e pretensiosa realizada em torres de marfim, certo? Se nenhuma pessoa comum consegue entender do que se trata, por que usar o bom dinheiro dos impostos para empregar gente para pensar nessas coisas? Vou inverter as coisas. Suponha que qualquer pessoa comum *pudesse* entender tudo o que os matemáticos pensam. Nesse caso, você ficaria feliz em usar o dinheiro dos impostos? Eles não estão sendo pagos pelo conhecimento específico que têm? Se tudo fosse tão fácil e compreensível que fizesse sentido imediato para qualquer um escolhido ao acaso na rua, de que adiantaria ter matemáticos? Se todo mundo soubesse drenar um sistema de aquecimento central e soldar uma junta, de que adiantaria ter encanadores?

Não posso lhe mostrar nenhum dispositivo sensacional que se baseie na conjectura de Hodge. Mas posso explicar sua importância dentro da matemática. A matemática moderna é um todo unificado, de modo que qualquer avanço significativo, em qualquer área central, acaba justificando seu valor em termos de dólares. Pode ser que não o tenhamos hoje na nossa cozinha, mas amanhã, quem sabe? Conceitos matemáticos estreitamente ligados já estão provando seu valor em diversas áreas da ciência, desde a física quântica e a teoria das cordas até robôs.

Às vezes aplicações práticas de matemática nova aparecem quase instantaneamente. Às vezes leva séculos. Nesse último caso, pode parecer mais efetivo em termos de custo aguardar até que surja a necessidade de tais resultados, e aí executar um programa de emergência para desenvolvê-los. Todos os problemas matemáticos que não tenham uso óbvio e imediato deveriam ser postos em banho-maria até que fossem necessários. No entanto, se fizéssemos isso sempre estaríamos correndo atrás, pois a matemática gastou algumas centenas de anos brincando de pega-pega com as necessidades da ciência aplicada. E pode ser que não fique muito clara a ideia da qual necessitamos. Você ficaria feliz se ninguém tivesse começado a pensar como se fazem tijolos até você ter de contratar um construtor para

começar a erguer uma casa? Quanto mais original é um conceito matemático, menos provável é que ele surja a partir de um programa de emergência.

Uma estratégia melhor é deixar algumas partes da matemática desenvolverem-se segundo suas próprias linhas, e deixar de esperar prova imediata. Não tente sempre ter vantagens imediatas; deixe que o edifício matemático cresça organicamente. Os matemáticos são baratos: não precisam de equipamento caro como os físicos de partículas (Grande Colisor de Hádrons: 7,5 bilhões de euros, por enquanto). Eles pagam suas despesas lecionando. Permita que alguns deles trabalhem em tempo parcial na conjectura de Hodge, se é isso que os atrai. Trata-se de algo bastante razoável.

VOU DESLINDAR A FORMULAÇÃO da conjectura de Hodge palavra por palavra. O conceito mais fácil é "variedade algébrica". É consequência natural do uso de coordenadas cartesianas ligar a geometria à álgebra (Capítulo 3). Com sua ajuda, o pequeno estojo de curvas introduzidas por Euclides e seus sucessores – reta, círculo, elipse, parábola, hipérbole – transformou-se numa cornucópia sem fundo. Uma reta, base da geometria euclidiana, é o conjunto de pontos que satisfazem uma equação algébrica característica, por exemplo, $y = 3x + 1$. Troque 3 e 1 por outros números, e você obterá outras retas. Os círculos necessitam de equações quadráticas, bem como as elipses, parábolas e hipérboles. Em princípio, qualquer coisa que se formule geometricamente pode ser reinterpretada algebricamente, e vice-versa. Então as coordenadas tornam a geometria obsoleta? Tornam a álgebra obsoleta? Por que usar duas ferramentas se uma faz o mesmo serviço que a outra?

No estojo de ferramentas que tenho na minha garagem, há um martelo e um grande par de alicates. O serviço do martelo é pregar pregos na madeira, o do alicate é puxá-los para fora. Em princípio, porém, eu poderia bater os pregos usando a lateral do alicate, e o martelo tem uma garra feita especificamente para extrair pregos. Então, para que preciso dos dois? Porque em alguns serviços o martelo é melhor, e em outros melhor é o alicate. O mesmo se passa com a geometria e a álgebra: algumas formas de pensar são mais naturais usando-se geometria, outras são mais naturais usando-se álgebra. É a ligação entre as duas que importa. Se o pensamento geométrico emperra, troque para a álgebra. Se o pensamento algébrico fica

emperrado, passe para a geometria.

A geometria de coordenadas oferece uma nova liberdade para inventar curvas. Basta escrever uma equação e buscar suas soluções. A menos que escolha uma equação boba como $x = x$, você deve obter uma curva. (A equação $x = x$ tem o plano inteiro como soluções.) Por exemplo, eu poderia escrever $x^3 + y^3 = 3xy$, cujas soluções estão desenhadas na Figura 45. Esta curva é o fólio de Descartes, e você não o achará em Euclides. A gama de novas curvas que qualquer um pode inventar é literalmente infinito.

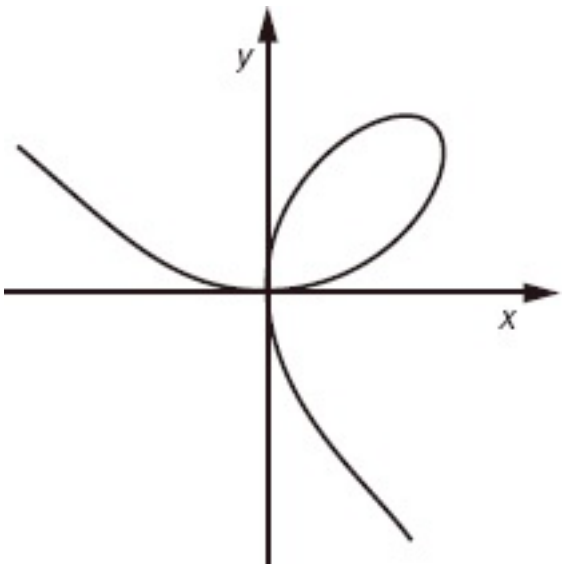


FIGURA 45 O fólio de Descartes.

Um reflexo automático entre os matemáticos é a generalização. Uma vez que alguém encontra uma ideia interessante, podemos perguntar se algo semelhante acontece num contexto mais geral. A ideia de Descartes tem pelo menos três generalizações ou modificações principais, todas elas necessárias para dar sentido à conjectura de Hodge.

Primeiro, o que acontece se trabalharmos com espaços diferentes do plano? O espaço euclidiano tridimensional tem três coordenadas (x, y, z) em vez de duas. No espaço, uma equação geralmente define uma superfície. Duas equações definem uma curva, onde as duas superfícies se encontram. Três equações geralmente determinam um ponto. (Por "geralmente" entendo que às vezes pode haver exceções, mas estas são muito incomuns e satisfazem condições especiais. Vimos algo similar no plano com a equação boba $x = x$.)

Mais uma vez, podemos definir novas superfícies ou curvas, não encontradas em Euclides, escrevendo novas equações. No século XIX foi

moda fazer isso. Você podia publicar uma nova superfície se dissesse algo genuinamente interessante sobre ela. Um exemplo típico é uma superfície introduzida por Kummer em 1864, com a equação

$$x^4 + y^4 + z^4 - y^2z^2 - z^2x^2 - x^2y^2 - x^2 - y^2 - z^2 + 1 = 0$$

A Figura 46 mostra uma imagem. As principais características de interesse são os dezesseis "pontos duplos", onde a forma é como dois cones grudados pela ponta. Este é o máximo número possível para uma superfície quártica, uma superfície de grau 4, e foi interessante o suficiente para merecer publicação.

No século XIX, os matemáticos haviam vivenciado os inebriantes deleites dos espaços de dimensões superiores. Não há necessidade de parar em três coordenadas; por que não experimentar quatro, cinco, seis … 1 milhão? Não é uma especulação de quem não tem o que fazer. É a álgebra de montes de equações em montes de variáveis, e elas aparecem em toda a paisagem matemática – por exemplo, no Capítulo 5 sobre a conjectura de Kepler e no Capítulo 8 sobre mecânica celeste. E tampouco uma generalização indolente: ser capaz de pensar em tais coisas geometricamente, bem como algebricamente, é uma ferramenta poderosa que não deveria ser restrita a espaços de duas ou três dimensões, só porque não é possível desenhar figuras e fazer modelos.

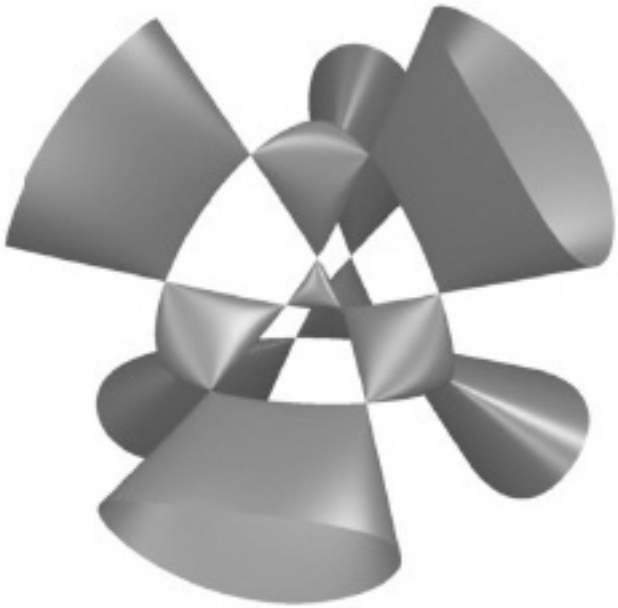


FIGURA 46 Superfície quártica de Kummer com seus dezesseis pontos duplos.

A palavra "dimensão" pode soar impressionante e mística, mas nesse

contexto tem um significado simples e direto: a quantidade de coordenadas de que você precisa. Por exemplo, o espaço 4-dimensional tem quatro coordenadas (*x*, *y*, *z*, *w*), e no que diz respeito à matemática, isso o define. Em quatro dimensões, uma equação única normalmente define uma "hipersuperfície" tridimensional, duas equações definem uma superfície (duas dimensões), três equações definem uma curva (uma dimensão) e quatro equações definem um ponto (zero dimensão). Cada nova equação libera uma dimensão – uma variável. Então, podemos predizer que no espaço de dezessete dimensões, onze equações definem um objeto 6-dimensional, exceto para raros (e detectáveis) casos onde algumas das equações são supérfluas.

Um objeto definido dessa maneira é chamado variedade algébrica, em inglês, *algebraic variety*. A palavra "*variety"* em inglês tem um sentido semelhante a "*manifold*". Por razões perdidas nas brumas da história, *manifold* ficou associada à topologia e à geometria diferencial (a topologia combinada com o cálculo), enquanto "*variety*" ficou associada à geometria algébrica.[1] O uso de palavras diferentes evita confusão, então em inglês os dois termos pegaram.[e] Uma variedade algébrica poderia ter sido chamada de "espaço multidimensional definido por um sistema de equações algébricas", mas é fácil de se ver por que ninguém fez isso.

Um segundo modo atraente de generalizar as noções da geometria analítica, a geometria de coordenadas, é permitir que as coordenadas sejam números complexos. Lembremos que o sistema de números complexos envolve um tipo novo de número, i, cujo quadrado é –1. Por que complicar tudo dessa maneira? Porque equações algébricas são muito mais bem-comportadas no domínio dos números complexos. Nos números reais, uma equação quadrática pode ter duas soluções ou nenhuma. (Pode ter também só uma, mas em um sentido significativo costuma-se dizer que é a mesma solução ocorrendo duas vezes.) Nos números complexos uma equação quadrática *sempre* tem duas soluções (mais uma vez, contando corretamente a multiplicidade). Para alguns propósitos, esta é uma propriedade muito mais agradável. Pode-se dizer "resolver a equação para a sétima variável" confiante de que tal solução efetivamente exista.

Por mais agradável que seja sob esse aspecto, a geometria algébrica complexa tem características que requerem que nos acostumemos um pouco. Com variáveis reais, uma reta pode cortar um círculo, ou ser tangente a ele, ou passar totalmente por fora. Com variáveis complexas, a terceira opção desaparece. Uma vez que nos acostumemos a essas

mudanças, porém, as variedades algébricas complexas são muito mais bem-comportadas do que as reais. Às vezes variáveis reais são essenciais, mas para a maioria dos propósitos o contexto complexo é uma escolha melhor. Em todo caso, agora sabemos o que é uma variedade algébrica complexa.

E quanto à "projetiva"? Esta é a terceira generalização, e requer uma noção de espaço ligeiramente diferente. A geometria projetiva surgiu a partir do interesse dos pintores da Renascença na perspectiva, e elimina o comportamento excepcional das retas paralelas. Em geometria euclidiana, duas retas ou se encontram ou são paralelas: não se encontram, por mais que sejam encompridadas. Agora, imagine-se de pé num plano infinito, pincel na mão, cavalete montado, paleta de tintas preparada, com um par de retas paralelas dirigindo-se para o pôr do sol distante como um par de trilhos infinitamente longos. O que você vê, e o que desenharia? Não duas linhas que nunca se encontram. Em vez disso, as linhas parecem convergir, encontrando-se no horizonte.

A que parte do plano o horizonte corresponde? À parte onde as retas paralelas se encontram. Mas isso não existe. O horizonte é o limite, na sua pintura, da imagem do plano. E se serve para o mundo, seguramente deveria ser a imagem do limite do plano. Mas um plano não tem limite. Ele continua para sempre. Tudo isso é um pouco intrigante. É como se parte do plano euclidiano estivesse faltando. Se você "projeta" um plano (aquele com os trilhos) em outro plano (a tela no cavalete), obtém uma reta na imagem, o horizonte, que não é a projeção de nenhuma linha no plano.

Existe um meio de se livrar dessa anomalia intrigante: acrescentar uma reta no infinito do plano euclidiano, representando o horizonte que falta. Agora tudo fica muito mais simples. Duas retas sempre se cruzam num ponto; a velha noção de retas paralelas corresponde ao caso em que duas retas se encontram no infinito. Essa ideia, adequadamente interpretada, pode ser transformada em matemática perfeitamente coerente. O resultado é chamado geometria projetiva. É um tema muito elegante, e os matemáticos dos séculos XVIII e XIX o adoravam. Eles acabaram esgotando o assunto, até que os matemáticos do século XX resolveram generalizar a geometria algébrica para espaços multidimensionais e usar números complexos. A essa altura ficou claro que poderíamos ocupar o terreno todo e estudar soluções complexas de sistemas de equações algébricas no espaço projetivo em vez de soluções reais no espaço euclidiano.

Deixe-me resumir. Uma variedade algébrica complexa projetiva é como

uma curva, definida por uma equação algébrica, mas:

- O número de equações e variáveis pode ser qualquer um que desejemos (variedade algébrica).
- As variáveis podem ser complexas em vez de reais (complexa).
- As variáveis podem assumir valores infinitos de forma coerente (projetiva).

E já que estamos aqui, há outro termo com que podemos lidar facilmente: não singular. Significa que a variedade é suave, sem protuberâncias ou locais onde a forma seja mais complicada do que uma suave região do espaço. A superfície de Kummer é singular naqueles dezesseis pontos duplos. É claro que temos de explicar o que significa "suave" quando as variáveis são complexas e algumas podem ser infinitas, mas essa é uma técnica de rotina.

ESTAMOS QUASE A MEIO CAMINHO da formulação da conjectura Hodge. Sabemos do que estamos falando, mas não como Hodge achou que ela devia se comportar. Agora precisamos lidar com os aspectos mais profundos e mais técnicos: ciclos algébricos, classes e (sobretudo) classes de Hodge. Contudo, posso revelar imediatamente a essência geral. São dispositivos técnicos que fornecem uma resposta parcial para uma pergunta muito básica acerca da nossa superfície generalizada: *qual é o formato dela?* Os únicos termos restantes, "combinação linear racional", fornecem o que todo mundo espera que seja a resposta correta para essa pergunta.

Veja até onde chegamos. Já compreendemos que tipo de afirmação é a conjectura de Hodge. Ela nos diz que dada qualquer superfície generalizada definida por algumas equações, pode-se determinar seu formato executando alguma álgebra com coisas chamadas ciclos. Eu poderia ter dito isso na primeira página deste capítulo, mas naquela etapa não teria feito mais sentido do que a afirmação formal que fiz. Agora que sabemos o que é uma variedade algébrica, tudo começa a se interligar.

Também começa a soar como topologia. "Achar o formato fazendo cálculos algébricos" é impressionantemente reminiscente das ideias de Poincaré a respeito dos invariantes algébricos para espaços topológicos. Então, o passo seguinte requer uma discussão de topologia algébrica. Entre

as descobertas de Poincaré havia três importantes tipos de invariante, definidos em termos de três conceitos: homotopia, homologia e co-homologia. A que queremos é a co-homologia – e, é claro, como você já deve bem ter desconfiado, essa é a mais difícil de explicar.

Acho que devemos simplesmente nos jogarmos nela.

No espaço tridimensional com coordenadas reais, uma esfera e um plano se cruzam (quando se cruzam) formando um círculo. A esfera é uma variedade algébrica, o círculo é uma variedade algébrica, e o círculo está contido na esfera. Nós o chamamos de *subvariedade* algébrica. Mais genericamente, se pegarmos as equações (muitas variáveis, complexas, projetivas) que definem alguma variedade algébrica, e acrescentarmos mais algumas equações, então tipicamente perdemos algumas das soluções: aquelas que deixam de satisfazer as novas equações. Quanto mais equações temos, menor se torna a variedade algébrica. O sistema extenso de equações define alguma parte da variedade algébrica original, e essa parte é por si só uma variedade algébrica – uma subvariedade.

Quando contamos o número de soluções de uma equação polinomial, pode ser conveniente contar o mesmo ponto mais de uma vez. Desse ponto de vista, o conjunto de soluções consiste de um número de pontos, sendo que em cada um "prendemos" um número, sua multiplicidade. Podemos, por exemplo, ter soluções 0, 1 e 2 com multiplicidades 3, 7 e 4, respectivamente. O polinômio então seria $x^3 (x - 1)^7 (x - 2)^4$, se você está interessado em saber. Cada um dos três pontos $x = 0$, 1 ou 2 é uma subvariedade (bastante trivial) dos números complexos. Então as soluções dessa equação polinomial podem ser descritas como uma lista de três subvariedades algébricas, com um número natural preso a cada uma, como um rótulo.

Um ciclo algébrico é similar. Em vez de pontos isolados, usamos qualquer lista finita de subvariedades. A cada uma delas podemos prender um rótulo numérico, que não precisa ser um número natural. Pode ser um inteiro negativo, pode ser um número fracionário racional, pode ser um número real não racional ou mesmo complexo. Por vários motivos, a conjectura de Hodge usa como rótulos números racionais. É a isso que se refere a "combinação linear racional". Assim, por exemplo, a nossa variedade algébrica original poderia ser a esfera unitária num espaço 11-dimensional, e esta lista poderia ter mais ou menos o seguinte aspecto:

Uma hiperesfera 7-dimensional (com equações tais e tais) com rótulo $^{22}/_{7}$.
Um toro (com equações tais e tais) com rótulo $-^{4}/_{5}$.
Uma curva (com equações tais e tais) com rótulo $^{413}/_{6}$.

Não tente visualizar isso, ou, se tentar, pense como um cartunista: três borrões rabiscados com pequenas legendas. Cada rabisco, cada lista, constitui um ciclo algébrico.

Por que fazer tanta confusão e se dar a todo esse trabalho para inventar algo tão abstrato? Porque ele capta aspectos essenciais da variedade algébrica original. Os geômetras algébricos estão tomando emprestado um truque dos topologistas.

NO CAPÍTULO 10, sobre a conjectura de Poincaré, pensamos numa formiga cujo universo é uma superfície. Como a formiga pode concluir de que formato é seu universo se não pode sair dele e dar uma olhada? Em particular, como ela pode distinguir uma esfera de um toro? A solução ali apresentada envolvia laçadas fechadas – viagens de ônibus topológicas. A formiga força essas laçadas, descobre o que acontece quando tem as pontas juntadas e calcula um invariante algébrico do espaço chamado de grupo fundamental. "Invariante" significa que espaços topologicamente equivalentes têm o mesmo grupo fundamental. Se os grupos são diferentes, então os espaços também são. Esse é o invariante que levou Poincaré à sua conjectura. No entanto, não é fácil para a pobre formiga examinar todas as laçadas possíveis em seu universo, e essa observação reflete sutilezas matemáticas genuínas no cálculo do grupo fundamental. Um invariante mais prático existe, e Poincaré também o investigou. Forçar laçadas é chamado de homotopia. Essa alternativa tem um nome similar: homologia.

Vou mostrar a versão mais simples, mais concreta, de homologia. Os topologistas rapidamente aperfeiçoaram essa versão, simplificaram, generalizaram e a transformaram numa imensa máquina matemática chamada álgebra homológica. Essa versão simples nos dá um tênue sabor de como funciona o tópico, mas não precisamos mais do que isso.

A formiga começa por fazer um levantamento do seu universo para criar um mapa. Como um agrimensor humano, ela cobre seu universo com uma rede de triângulos. A condição crucial é que triângulo algum esteja em torno de um furo na superfície, e o meio de assegurar isso é fazer os

triângulos grudando retalhos de borracha na superfície, como alguém consertando um pneu de bicicleta. Assim, cada triângulo tem um interior bem-definido que é topologicamente o mesmo que o interior de um triângulo comum no plano. Topologistas chamam esse retalho de disco topológico, porque topologicamente ele é equivalente a um disco e seu interior. Para ver por quê, olhe a Figura 36 no Capítulo 10, onde o triângulo é deformado continuamente até se tornar um círculo. Não é possível grudar um retalho desse tipo num triângulo que cerca um furo porque o furo cria um túnel que liga o interior do triângulo ao exterior. O retalho teria que abandonar a superfície, e a formiga não tem permissão para fazer isso.

A formiga criou agora uma *triangulação* em seu universo. A condição referente aos retalhos assegurou que a topologia da superfície – seu formato, no sentido de equivalência topológica – possa ser reconstituída se tudo que soubermos for a lista de triângulos, junto com a informação de quais triângulos são adjacentes a quais. Se você fosse a uma loja de móveis modulares e comprasse um kit de triângulos convenientemente rotulados, e aí colasse a borda A com a borda AA, a borda B com a BB, e assim por diante, seria capaz de construir a superfície. A formiga está confinada à superfície, então não pode fazer um modelo, mas pode ter certeza de que em princípio seu mapa contém a informação de que necessita. Para extrair essa informação, ela precisa realizar um cálculo. Ao fazê-lo, a formiga não precisa mais contemplar a infinidade de todas as laçadas possíveis, mas precisa, sim, contemplar um bocado delas: todas as laçadas fechadas que correm ao longo das bordas de sua rede escolhida.

Em homotopia, perguntamos se uma dada laçada pode ser encolhida continuamente até virar um ponto. Em homologia, fazemos uma pergunta diferente: a laçada forma a fronteira de um disco topológico? Ou seja, pode-se encaixar um ou mais retalhos triangulares de modo que o resultado seja uma região sem nenhum furo e a fronteira dessa região seja a referida laçada?

A Figura 47 (esquerda) mostra parte da triangulação de uma esfera, uma laçada fechada, e o disco topológico cuja fronteira é essa laçada. Escolhendo as técnicas certas, pode-se provar que *qualquer* laçada numa triangulação de uma esfera é uma fronteira: retalhos triangulares, e mais genericamente discos topológicos, são detectores de furos, e intuitivamente uma esfera não tem furos. Entretanto, o toro tem um furo, e de fato algumas laçadas num toro não são fronteiras. A Figura 47 (direita) mostra uma laçada dessas, percorrendo o furo central. Em outras palavras: ao examinar

uma lista de laçadas e descobrir quais delas são fronteiras, a formiga pode distinguir um universo esférico de um universo toroidal.

Se for uma formiga esperta, como Poincaré e outros topologistas de sua época, ela pode transformar essa ideia num elegante invariante topológico, o grupo de homologia de sua superfície. A ideia básica é "somar" duas laçadas desenhando ambas. Contudo, isso não é uma laçada, então temos que voltar e começar de novo. Na verdade, retornar até o início de tudo; voltar aos dias em que fomos apresentados pela primeira vez à álgebra. Minha professora de matemática começou mostrando que você pode somar um número de maçãs com um número de maçãs e obter um número total de maçãs. Mas não pode somar maçãs com laranjas, a não ser que conte tudo como frutas.

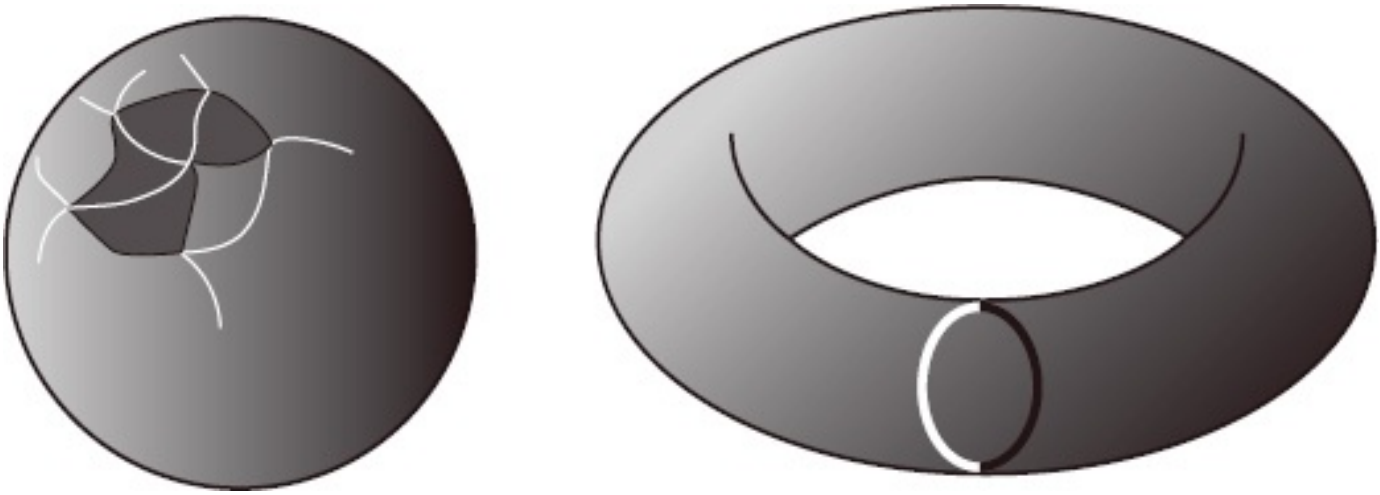


FIGURA 47 *Esquerda*: Parte da triangulação de uma esfera, uma laçada fechada (linhas pretas), e o disco cuja fronteira é essa laçada (sombreamento escuro). *Direita*: Laçada num toro que não é a fronteira de um disco (a parte mais clara está atrás).

ISSO É VERDADE EM ARITMÉTICA, embora mesmo aí você precise ter cuidado para não usar a mesma maçã duas vezes, mas não é verdade em álgebra. Aí, você pode somar maçãs com laranjas, ao mesmo tempo em que as mantêm distintas. Na verdade, em matemática avançada, é lugar-comum somar coisas que se poderia imaginar que ninguém em seu juízo perfeito teria pensado em inventar, muito menos somar umas com as outras. A liberdade para fazer esse tipo de coisa acaba se revelando tremendamente útil e importante, e os matemáticos que o fizeram afinal não eram loucos – pelo menos, não sob este aspecto.

Para entender algumas das ideias que a conjectura de Hodge reúne, precisamos ser capazes de somar maçãs com laranjas sem agrupá-las todas

como simples frutas. O modo de somá-las não é realmente muito difícil. O que é difícil é aceitar que existe sentido em fazê-lo. Muitos de nós já nos deparamos com uma versão desse bloqueio conceitual em potencial. Minha professora disse à classe que as letras representavam números desconhecidos, com letras diferentes para números diferentes. Se você tivesse a maçãs e outras a maçãs, o número total de maçãs era $a + a = 2a$. E isso funcionava qualquer que fosse o número de maçãs. Se você tivesse $3a$ maçãs e somasse $2a$ maçãs, o resultado era $5a$ maçãs, qualquer que fosse o número de maçãs. O símbolo, e o que ele representava, tampouco tinha importância: se você tivesse $3b$ laranjas e somasse $2b$ laranjas, o resultado era $5b$ laranjas.[2] Mas o que acontecia quando você tinha $3a$ maçãs e $2b$ laranjas? Quanto era $3a + 2b$?

$3a + 2b$.

Era isso aí. Não se podia simplificar a soma e dar 5 algumas coisas – pelo menos não sem algumas manipulações envolvendo uma categoria nova, frutas, e algumas novas equações. Era o melhor que se podia fazer: conviver com isso. No entanto, uma vez dado esse passo, você podia fazer somas do tipo

$$(3a + 2b) + (5a - b) = 8a + b$$

sem introduzir ideias novas. Ou novos tipos de fruta.

Havia alguns empecilhos. Já comentei que se você soma uma maçã com uma maçã, você apenas obtém duas maçãs se a segunda for diferente da primeira. A mesma coisa vale para combinações mais complicadas de maçãs e laranjas. A álgebra pressupõe que, com o propósito de fazer somas, todas as maçãs envolvidas sejam diferentes. Na verdade, geralmente faz sentido fazer essa suposição, mesmo em casos nos quais duas maçãs – ou qualquer outra coisa que estejamos somando – possam na realidade ser a mesma. Uma maçã mais a mesma maçã é uma maçã com multiplicidade dois.

Uma vez que você se acostuma com essa ideia, pode usá-la para qualquer coisa. Um porco mais o mesmo porco é esse porco com multiplicidade dois: porco + porco = 2 porcos, seja lá que porco for. Um porco mais uma vaca é porco + vaca. Um triângulo mais três círculos é triângulo + 3 círculos. Uma superduperesfera mais três quasipilhas hiperlípticas é

superduperesfera + 3 quasipilhas hiperlípticas

o que quer que signifique a nomenclatura (que, aliás, não significa nada).

Você pode até permitir números negativos, e falar de três porcos menos onze vacas: 3 porcos −11 vacas. Não tenho a menor ideia do aspecto de −11 vacas, mas posso estar confiante de que se eu somar seis vacas a isso, obterei menos cinco vacas.[3] É um jogo formal jogado com símbolos, e nenhuma interpretação mais realista é necessária, ou – frequentemente – possível. Você pode permitir números reais: π porcos menos $\sqrt{2}$ vacas. E números complexos. Qualquer tipo de número fantástico que algum matemático já tenha inventado ou inventará no futuro. A ideia pode se tornar um pouco mais respeitável se você pensar nos números como *rótulos*, presos aos porcos e vacas. Agora π porcos menos $\sqrt{2}$ vacas pode ser pensado como um porco com rótulo π junto com uma vaca rotulada $-\sqrt{2}$. A aritmética aplica-se aos rótulos, não aos animais.

A conjectura de Hodge envolve esse tipo de construção, com sinos e apitos extras. Em lugar de animais, usa curvas, superfícies e suas análogas de dimensões superiores. Por estranho que possa parecer, o resultado não é um simples absurdo abstrato, mas uma profunda conexão entre topologia, álgebra, geometria e análise.

PARA ESTABELECER O FORMALISMO de homologia queremos somar laçadas, mas não como fizemos para o grupo fundamental. Em vez disso, o fazemos do modo como a minha professora ensinou. Simplesmente anotamos as laçadas e pomos um sinal de + no meio. Para dar sentido a isso, trabalhamos não com laçadas isoladas, mas com conjuntos finitos de laçadas. Rotulamos cada laçada com um número natural que conta quantas vezes ela aparece. Chamemos esse conjunto rotulado de *ciclo*. Agora a formiga pode somar quaisquer dois ciclos agrupando-os e somando os rótulos correspondentes, e o resultado é outro ciclo. Talvez eu devesse ter usado bicicletas em vez de ônibus, para minha imagem das viagens da formiga no Capítulo 10.

Quando estávamos construindo o grupo fundamental, onde "somar" laçadas pelas extremidades, surgiu um nó técnico. Somar a laçada trivial a uma laçada não resultava *exatamente* na mesma laçada, então a laçada zero era malcomportada. Somar uma laçada à sua oposta não resultava

exatamente na laçada trivial, então as opostas não se comportavam corretamente. A saída foi considerar as laçadas como sendo a mesma se uma pudesse ser deformada na outra.

Para a homologia, esse não é o problema. Existe um ciclo zero (todos os rótulos zero), e todo ciclo tem um oposto (transforma todo rótulo em seu negativo), então obtemos sim um grupo. O problema é que… bem, é o grupo errado. Ele não nos diz nada a respeito da topologia do espaço. Para resolver isso usamos um artifício semelhante, adotando uma visão mais relaxada de que ciclos devem contar como zero. A formiga corta o espaço em retalhos triangulares, e a fronteira de cada retalho é topologicamente bastante trivial: você pode encolhê-la até virar um ponto, forçando tudo até o meio do retalho. Então fazemos com que esses ciclos de fronteira sejam equivalentes ao ciclo zero. É um pouco como transformar números comuns em aritmética de relógio, fingindo que o número doze é irrelevante, de modo que pode virar zero. Aqui transformamos ciclos em homologia fingindo que qualquer ciclo de fronteira é irrelevante.

As consequências dessa pretensão são dramáticas. Agora a álgebra dos ciclos é afetada pela topologia do espaço. O grupo de ciclos módulo fronteiras é um invariante topológico útil, o grupo de homologia da superfície. À primeira vista ele depende de qual triangulação a formiga escolhe, mas, devido à característica de Euler, diferentes triangulações da mesma superfície conduzem ao mesmo grupo de homologia. Assim, a formiga inventou um invariante algébrico capaz de distinguir superfícies diferentes. É um pouco trabalhoso, mas nunca se obtém bons invariantes sem realizar algum trabalho pesado em algum ponto do caminho. Esse aqui é tão efetivo que consegue distinguir não só uma esfera de um toro, mas um toro de dois furos de um toro de cinco furos, e da mesma maneira para outras quantidades de furos.

A HOMOLOGIA PODE PARECER um pouco pretensiosa, mas deu início a um rico veio de invariantes topológicos, e baseia-se em ideias geométricas simples: laçadas, fronteiras, agrupamento de conjuntos, realizar aritmética com rótulos. Considerando que a pobre formiga está confinada à sua superfície, é estarrecedor que aquela criatura possa descobrir qualquer coisa significativa sobre o formato de seu universo simplesmente grudando retalhos triangulares, criando um mapa e fazendo um pouco de álgebra.

Existe um modo natural de estender a homologia a dimensões

superiores. O análogo 3-dimensional de um triângulo é o tetraedro; ele tem quatro vértices, seis arestas, quatro faces triangulares e uma única "face" 3-dimensional, seu interior. Mais genericamente, em *n* dimensões podemos definir um *n*-simplex com *n* + 1 vértices, ligados em pares por todas as arestas possíveis, que por sua vez formam triângulos, que se juntam para criar tetraedros, e assim por diante. Agora é fácil definir ciclos, fronteiras e homologia. E mais uma vez podemos conceber um grupo somando (classe de homologia de) ciclos. Na verdade, obtemos agora toda uma série de grupos: um para ciclos 0-dimensionais (pontos), um para ciclos 1-dimensionais (retas), um para ciclos 2-dimensionais (triângulos), e assim por diante, até chegarmos à dimensão do próprio espaço. Estes são o 0º, 1º, 2º, e assim por diante, grupos de homologia do espaço. Grosso modo, eles dão a precisa noção dos furos, das várias dimensões, no espaço: existem? Quantos são? E como se relacionam entre si?

Então, isso é homologia, e é quase o que precisamos para entender o que *diz* a conjectura de Hodge. Porém, o que realmente precisamos é de um conceito estreitamente ligado chamado *co-homologia*. Em 1893, Poincaré notou uma curiosa coincidência na homologia de qualquer variedade:[f] a lista de grupos de homologia é igual, porém invertida. Para uma variedade de dimensão 5, digamos, o 0º grupo de homologia é o mesmo que o 5º, o 1º grupo de homologia é igual ao 4º, e o 2º grupo de homologia é o mesmo que o 3º. Ele percebeu que isso não podia ser mera coincidência, e explicou o fato em termos da triangulação dual, com a qual travamos conhecimento no Capítulo 4 em relação aos mapas. Esta é uma segunda triangulação na qual cada triângulo é substituído por um vértice, cada fronteira entre dois triângulos por um lado ligando os dois novos vértices correspondentes, e cada ponto por um triângulo, como na Figura 9 do Capítulo 4. Note como as dimensões aparecem em ordem inversa: triângulos 2-dimensionais tornam-se pontos 0-dimensionais, e vice-versa; lados 1-dimensionais permanecem 1-dimensionais porque 1 está exatamente no meio.

Acaba-se descobrindo que vale a pena distinguir as duas listas, mesmo que elas produzam os mesmos invariantes. Quando o processo todo é generalizado e formulado em termos abstratos, as triangulações desaparecem, e a triangulação dual deixa de fazer sentido. O que sobrevive são duas séries de invariantes topológicos, chamadas grupos de homologia e grupos de co-homologia. Todo conceito em homologia tem um dual, geralmente batizado adicionando-se o prefixo co-. Logo, em lugar de ciclos temos cociclos, e em lugar de dois ciclos serem homólogos temos dois

cociclos sendo co-homólogos. As classes a que se refere a conjectura de Hodge são classes co-homólogas, e estas são coleções de cociclos co-homólogos entre si.

Homologia e co-homologia não nos contam tudo que gostaríamos de saber sobre o formato de um espaço topológico – espaços distintos podem ter a mesma homologia e co-homologia –, mas fornecem, sim, um bocado de informação útil e um arcabouço sistemático para calcular e usar.

UMA VARIEDADE ALGÉBRICA, seja real, complexa, projetiva ou não, é um espaço topológico. Portanto tem uma forma. Para descobrir coisas úteis sobre essa forma, pensamos como topologistas e calculamos os grupos de homologia e co-homologia. Mas os ingredientes naturais na geometria algébrica não são objetos geométricos como triangulações e ciclos. São coisas que podemos descrever com maior facilidade como equações algébricas. Volte e olhe a equação da superfície de Kummer. Como ela se relacionaria com triangulação? Não existe nada na fórmula que dê o menor indício de triângulos.

Talvez tenhamos que começar de novo. Em vez de triângulos, deveríamos usar os blocos construtivos naturais para variedades algébricas, que são as subvariedades, definidas impondo-se equações adicionais. Agora precisamos redefinir ciclos: em vez de conjuntos de triângulos com rótulos de números naturais, usamos conjuntos de subvariedades com o rótulo que melhor sirva para elas. Por vários motivos – sobretudo pelo fato de que a conjectura de Hodge é falsa se usarmos rótulos com números inteiros naturais – os números racionais são a escolha sensata. A pergunta de Hodge fica então reduzida ao seguinte: será que essa nova definição de homologia e co-homologia capta tudo aquilo que a definição topológica capta? Se a conjectura for verdadeira, então a ferramenta do ciclo algébrico é bastante afiada para servir como cinzel co-homológico da topologia. Se for falsa, então o ciclo algébrico é um instrumento cego, sem fio.

Exceto… Perdão, bati demais o pudim e ele desandou. A conjectura diz que basta usar um *tipo* particular de ciclo algébrico, um ciclo que viva numa classe de Hodge. Para explicar isso, precisamos de mais um ingrediente nessa já rica mistura: a análise. Um dos conceitos mais importantes em análise é o da equação diferencial, que é uma condição acerca das taxas de variação das variáveis (Capítulo 8). Quase toda a física matemática dos

séculos XVIII, XIX e XX modela a natureza usando equações diferenciais, e o mesmo ainda ocorre no século XXI. Nos anos 1930, essa ideia levou Hodge a um novo corpo de técnica, agora chamado teoria de Hodge. Este corpo liga-se naturalmente a uma porção de outros métodos poderosos na área geral da análise e da topologia.

A ideia de Hodge foi usar uma equação diferencial para organizar as classes de co-homologia em distintos tipos. Cada peça tem uma estrutura extra, que pode ser utilizada com proveito em problemas topológicos. As peças são definidas usando-se uma equação diferencial que surgiu no final dos anos 1700, notavelmente no trabalho de Pierre-Simon de Laplace. É chamada, de modo coerente, de equação de Laplace. A principal pesquisa de Laplace era em mecânica celeste, o movimento e a forma de planetas, luas, cometas e estrelas. Em 1783, estava trabalhando no formato detalhado da Terra. A essa altura, sabia-se que a Terra não era esférica, mas achatada nos polos de modo a formar um esferoide oblato – como uma bola de praia com alguém sentado em cima. Porém, mesmo essa descrição deixa de fora alguns detalhes sutis. Laplace achou um método para calcular o formato com qualquer precisão que se desejasse, baseado em uma grandeza física que representa o campo gravitacional da Terra: não o campo em si, mas um potencial gravitacional. Este é uma medida da energia contida na gravitação, uma grandeza numérica definida em cada ponto do espaço. A força da gravidade atua na direção que faça o potencial decrescer mais rápido, e a magnitude da força é a taxa de decréscimo.

O potencial satisfaz a equação de Laplace: grosso modo, ela diz que na ausência de matéria – ou seja, no vácuo – o valor médio do potencial sobre uma esfera muito pequena é o mesmo que o valor no centro da esfera. É uma espécie de democracia: o seu valor é a média dos valores dos seus vizinhos. Qualquer solução da equação de Laplace é chamada função harmônica. Os tipos especiais de classe de co-homologia de Hodge são aqueles que possuem uma relação particular com funções harmônicas. A teoria de Hodge, o estudo desses tipos, inaugurou uma profunda e magnífica área da matemática: relações entre a topologia do espaço e uma equação diferencial especial sobre esse espaço.

Então, agora aí está. A conjectura de Hodge postula uma conexão profunda e poderosa entre três dos pilares da matemática moderna: álgebra, topologia e análise. Peguemos uma variedade algébrica qualquer. Para entender seu formato (topologia, levando a classes de co-homologia) escolha casos especiais delas (análise, levando a classes de Hodge mediante

equações diferenciais). Esses tipos especiais de classes de co-homologia podem ser entendidos usando-se subvariedades (álgebra: introduza algumas equações adicionais e olhe os ciclos algébricos). Isto é, para solucionar o problema de topologia "qual é o formato desta coisa?" para uma variedade algébrica, transforma-se a questão em análise e aí resolve-se usando álgebra.

Por que isso é importante? A conjectura de Hodge é uma proposta de acrescentar duas novas ferramentas ao estojo do geômetra algébrico: invariantes topológicos e equação de Laplace. Não é realmente uma conjectura a respeito de um teorema matemático; é uma conjectura sobre novos tipos de ferramentas. Se esta conjectura for verdadeira, essas ferramentas adquirem uma nova significação, e podem ser potencialmente usadas para responder a uma corrente infindável de questões. É claro que ela pode acabar se revelando falsa. Isso seria decepcionante, mas é melhor entender as limitações de uma ferramenta do que continuar martelando o dedo com ela.

AGORA QUE APRECIAMOS a natureza da conjectura de Hodge, podemos buscar evidência para ela. O que sabemos? Muito pouco.

Em 1924, antes de Hodge fazer sua conjectura, Solomon Lefschetz provou um teorema que recai na conjectura de Hodge para a co-homologia de dimensão-2 de qualquer variedade algébrica. Com um pouco de topologia algébrica rotineira, isso confirma a conjectura de Hodge para variedades algébricas de dimensões 1, 2 e 3. Para variedades algébricas de dimensões superiores, são conhecidos apenas alguns poucos casos especiais da conjectura de Hodge.

Hodge originalmente formulou sua conjectura em termos de rótulos com números naturais. Em 1961, Michael Atiyah e Friedrich Hirzebruch provaram que em dimensões superiores essa versão da conjectura é falsa. Assim, hoje interpretamos a conjectura de Hodge usando rótulos racionais. Para essa versão, há uma certa quantidade de evidências encorajadoras. A evidência mais forte em seu favor é que uma de suas consequências mais profundas, um teorema ainda mais técnico, conhecido como "algebricidade dos loci de Hodge", foi provado – *sem* pressupor a conjectura de Hodge. Eduardo Cattani, Pierre Deligne e Aroldo Kaplan acharam essa prova em 1995.

Finalmente, há uma atraente conjectura em teoria dos números que é

análoga à conjectura de Hodge. É chamada conjectura de Tate, em nome de John Tate, e liga a geometria algébrica à teoria de Galois, o círculo de ideias que provou que não existe fórmula algébrica para resolver equações polinomiais de grau 5. Sua formulação é técnica, e envolve ainda outra versão de co-homologia. Há razões independentes para ter esperança de que a conjectura de Tate possa ser verdadeira, mas seu status hoje está em aberto. Porém ao menos existe uma relação sensível com a conjectura de Hodge, mesmo que por enquanto pareça ser igualmente intratável.

A conjectura de Hodge é uma dessas incômodas asserções matemáticas para a qual a evidência a favor ou contra não é muito extensiva, nem particularmente convincente. Existe, sem dúvida, perigo de que a conjectura possa estar errada. Talvez haja uma variedade algébrica com 1 milhão de dimensões que refute a conjectura de Hodge, por razões que se concentrem em uma série de cálculos sem estrutura, tão complicados que ninguém jamais os realize. Se for assim, a conjectura de Hodge poderia ser falsa por motivos essencialmente tolos – apenas por não ser verdadeira –, mas de fato impossível de provar. Conheço alguns geômetras algébricos que desconfiam ser esse o caso. Então, esse milhão de dólares ficará a salvo para um futuro próximo.

---

[e] O mesmo não se pode dizer em português. *"Manifold"* é "variedade", conceito em topologia e geometria diferencial; *"algebraic variety"* é "variedade algébrica", portanto em português precisamos ser cuidadosos para evitar confusão. (N.T.)

[f] *Manifold* – portanto, no sentido topológico, e não de variedade algébrica. (N.T.)

16. E agora, para onde?

PREVER É ALGO MUITO DIFÍCIL, sobretudo o futuro,[1] como se supõe terem dito o físico ganhador do Nobel Niels Bohr e o jogador e técnico de beisebol Yogi Berra.[2] Aliás, Berra também falou: "Eu nunca disse a maioria das coisas que eu disse." Alegadamente. Arthur C. Clarke, famoso pela sua ficção científica e pelo filme *2001: Uma odisseia no espaço* e suas continuações, também foi futurologista: escreveu livros predizendo o futuro da tecnologia e da sociedade. Entre as muitas predições de seu *Perfil do futuro* estão:

Compreensão da linguagem das baleias e dos golfinhos em 1970.
Energia de fusão em 1990.
Detecção de ondas de gravidade em 1990.
Colonização de planetas em 2000.

Nada disso ainda aconteceu. Por outro lado, ele teve alguns sucessos:

Pousos planetários em 1980 (embora talvez tenha se referido a pousos humanos).
Máquinas tradutoras em 1970 (um pouco prematuro, mas agora existem no Google).
Rádio pessoal em 1990 (telefones celulares funcionam dessa maneira).

Ele predisse também que teríamos uma biblioteca global por volta de 2000, e este pode ser o alvo mais certeiro do que imaginávamos alguns anos atrás, pois esta é uma das muitas funções da internet. Com o advento da computação de nuvem, podemos todos acabar utilizando o mesmo computador gigante. Ele deixou de ver algumas das tendências mais importantes, como o surgimento do computador pessoal e a engenharia genética, embora a tenha predito para 2030. Com os resultados desiguais de Clarke como alerta, seria tolice predizer em detalhe o futuro dos grandes problemas matemáticos. No entanto, posso fazer algumas suposições, certo de que a maioria delas vai dar errado.

Na introdução mencionei a lista feita por Hilbert em 1900 dos 23 grandes problemas. A maioria deles está agora resolvida, e o seu corajoso grito de guerra "Nós precisamos saber, nós havemos de saber" parece agora justificado. Todavia, ele disse também "em matemática não há *ignorabimus* [seremos ignorantes]" e Kurt Gödel contradisse essa ideia com seu teorema da incompletude: alguns problemas matemáticos podem não ter solução dentro do arcabouço lógico da matemática. Não só impossíveis, como a quadratura do círculo: podem ser indecidíveis, significando que não existe prova nem refutação. É possível que possa ser esta a sorte de alguns dos grandes problemas até hoje não resolvidos. Eu ficaria surpreso caso a hipótese de Riemann fosse assim, e perplexo se alguém pudesse prová-la indecidível, mesmo que o fosse. Por outro lado, o problema P/NP pode muito bem ser indecidível, ou satisfazer alguma outra variação técnica do tema "não pode ser feito". Bem … Pelo menos é o que deixa transparecer.

Desconfio que por volta do final do século XXI teremos provas da hipótese de Riemann, da conjectura de Birch–Swinnerton-Dyer e da hipótese do *mass gap*, junto com refutações da conjectura de Hodge e da regularidade de soluções da equação de Navier-Stokes em três dimensões. Acredito que P/NP ainda não estará solucionado em 2100, mas sucumbirá em algum momento no século XXII. Então é claro que alguém refutará a hipótese de Riemann amanhã e provará que P é diferente de NP na semana que vem.

Eu estou em terreno mais seguro com observações gerais, porque podemos aprender com a história. Assim, estou razoavelmente confiante de que na época em que os sete problemas do milênio estiverem solucionados, muitos parecerão curiosidades históricas menores. "Ah, costumavam achar que *isso* era importante, não é?" Foi isso o que aconteceu com alguns dos problemas da parada de sucessos de Hilbert. Posso estar confiante também de que dentro de cinquenta anos diversas áreas fundamentais na matemática que não existem hoje vão passar a existir. E aí transpirará que alguns conceitos básicos e alguns teoremas rudimentares nessas áreas já existiam há muito tempo, mas ninguém percebeu que esses fragmentos isolados eram pistas para novas áreas profundas e importantes. Foi isso o que aconteceu com a teoria de grupos, a álgebra matricial, os fractais e o caos. Não duvido que acontecerá de novo, porque é um dos padrões segundo os quais a matemática evolui.

Essas novas áreas surgirão por meio de dois fatores principais. Emergirão a partir da estrutura interna da própria matemática ou virão

como respostas a novas questões sobre o mundo exterior – frequentemente as duas coisas em sintonia. Como o processo de três passos de Poincaré para solução de problemas – preparação, incubação e iluminação –, a relação entre a matemática e suas aplicações não é uma transição única: a ciência apresenta um problema, a matemática o resolve, e pronto. Em vez disso, encontramos uma intrincada rede de trocas de questões e ideias, com a matemática provocando novos experimentos, observações ou teorias, que por sua vez motivam uma nova matemática. E cada nó dessa rede, descobrimos ao examinar de perto, revela-se uma rede menor do mesmo tipo.

Existe mais mundo exterior do que costumava existir. Até recentemente, a principal força de inspiração externa para a matemática eram as ciências físicas. Algumas outras áreas desempenharam seu papel: biologia e sociologia influenciaram o desenvolvimento da probabilidade e da estatística, e a filosofia teve um grande efeito na lógica matemática. No futuro veremos contribuições crescentes da biologia, medicina, computação, finanças, economia, sociologia e muito possivelmente política, indústria cinematográfica e esportes. Desconfio que alguns dos primeiros novos grandes problemas surgirão da biologia, porque esse elo está agora firme no lugar. Uma tendência é a nossa capacidade de reunir dados biológicos e bioquímicos; pequenos genomas podem agora ser sequenciados usando um dispositivo do tamanho de um microcartão de memória, baseado em tecnologia de nanoporos, por exemplo. Grandes genomas rapidamente seguirão o mesmo caminho, usando esta ou outra tecnologia, a maior parte já existente.

Esses desenvolvimentos são potencialmente capazes de mudar o jogo, mas precisamos ter métodos melhores para compreender a implicação dos dados. A biologia não trata realmente dos dados em si. Ela trata de processos. A evolução é um processo, e também o são a divisão celular, o crescimento do embrião, a instalação de um câncer, o movimento das massas, o funcionamento do cérebro e a dinâmica do ecossistema global. A melhor maneira de conhecermos atualmente os ingredientes básicos de um processo, e deduzir o que ele faz, é a matemática. Assim, haverá grandes problemas de novos tipos – como a dinâmica se desenrola na presença de informação organizadora complexa, mas específica (sequências de DNA); como mudanças genéticas conspiram com o meio ambiente para restringir a evolução; como regras para crescimento, divisão, mobilidade, aderência e morte de células dão a organismos em desenvolvimento sua forma; como o

fluxo de elétrons e componentes químicos numa rede de células nervosas determina o que ela pode perceber e como agirá.

Computação é outra fonte de matemática nova que já tem histórico. Ela geralmente é considerada uma ferramenta para se fazer matemática, mas a matemática é também uma ferramenta para compreender e estruturar a computação. Essa via de mão dupla está se tornando cada vez mais importante para a saúde e o desenvolvimento de ambas as áreas, e é possível até que em algum ponto no futuro elas se fundam. Alguns matemáticos sentem que jamais deveriam ter permitido a cisão. Entre os muitos caminhos aqui visíveis, salta mais uma vez à mente a questão de conjuntos de dados muito grandes. Ela relaciona-se não só com o exemplo de DNA mencionado antes, mas com a predição de terremotos, evolução, clima global, mercado de ações, finanças internacionais e novas tecnologias. O problema é usar grandes quantidades de dados para testar e refinar modelos matemáticos do mundo real, de modo que nos deem um controle genuíno sobre sistemas muito complexos.

A predição sobre a qual estou mais confiante é, sob alguns aspectos, negativa, mas também uma afirmação da contínua criatividade da comunidade matemática. Todos os matemáticos de pesquisa sentem, de tempos em tempos, que seu tema tem uma mente própria. Os problemas se resolvem da maneira como a matemática quer que se resolvam, e não como os matemáticos querem. Podemos escolher que perguntas fazer, mas não podemos escolher as respostas que obtemos. Essa sensação está relacionada com duas importantes escolas de pensamento sobre a natureza da matemática. Os platônicos pensam que as "formas ideais" da matemática possuem algum tipo de existência independente, "lá fora", em algum reino distinto do mundo físico. (Há formas mais sutis de se dizer isso, formas que soam provavelmente mais coerentes, mas a essência é essa.) Outros veem a matemática como um constructo humano partilhado. Porém, ao contrário da maioria das coisas – o sistema legal, dinheiro, ética, moralidade –, a matemática é um constructo com um forte esqueleto lógico. Há severas restrições sobre o tipo de afirmações que se pode ou não se pode compartilhar com os outros. São essas restrições que dão a impressão de que a matemática tem seu próprio programa, e criam a sensação na mente do matemático de que a matemática em si *existe* fora do domínio da atividade humana. O platonismo, penso eu, não é uma descrição do que a matemática é. É uma descrição da *sensação* que a matemática dá quando se trabalha com ela. É como a vívida sensação de "vermelho" que

experimentamos quando vemos uma rosa, sangue ou um farol de trânsito. Os filósofos chamam essas sensações de qualia (singular: quale), e alguns deles pensam que a nossa sensação de livre-arbítrio é, na verdade, uma quale do modo do cérebro de tomar decisões. Quando decidimos entre alternativas, sentimos que temos uma escolha genuína – seja ou não a dinâmica do cérebro realmente determinística em algum sentido. De modo semelhante, o platonismo é uma quale de tomar parte num constructo humano partilhado dentro de um arcabouço rígido de dedução lógica.

Então a matemática pode parecer ter mente própria, mesmo que seja uma criação coletiva de mentes humanas. A história nos diz que a mente matemática – neste sentido – é mais inovadora e surpreendente do que qualquer mente humana individual pode prever. Tudo isso é um modo complicado de chegar ao meu ponto principal: uma coisa que podemos prever com segurança sobre o futuro da matemática é que ele será imprevisível. As questões matemáticas mais importantes do próximo século emergirão como consequências naturais, até mesmo inevitáveis, da nossa crescente compreensão do que atualmente acreditamos ser os grandes problemas da matemática. No entanto, é quase certo que serão questões que hoje nem podemos conceber. Isto é certo e aceitável, e devemos celebrar.

17. Doze para o futuro

NÃO QUERO DEIXAR você com a impressão de que a maioria dos problemas matemáticos foi resolvida, exceto por aquele esquisito realmente difícil. A pesquisa matemática é como explorar um novo continente. À medida que a área conhecida vai se expandindo, a fronteira que beira o desconhecido se amplia. Não estou sugerindo que quanto mais matemática descobrimos, menos sabemos; estou dizendo que quanto mais matemática descobrimos, mais percebemos que não sabemos. Mas o que não sabemos muda com o passar do tempo, com alguns problemas antigos desaparecendo enquanto novos são acrescentados. Em contraste, o que sabemos simplesmente aumenta – barrando o eventual documento perdido.

Para dar um pequeno indício do que atualmente *não* sabemos, além dos grandes problemas já discutidos, eis doze problemas não solucionados que vêm atormentando o mundo dos matemáticos há algum tempo. Eu os escolhi de modo que seja fácil entender as questões. Como foi amplamente demonstrado, isso não traz indicação alguma de quanto pode ser fácil ou difícil encontrar as respostas. Alguns deles podem acabar se revelando grandes problemas: isso dependerá sobretudo dos métodos inventados para resolvê-los e a que eles nos conduzem, não da resposta em si.

Problema de Brocard

Para qualquer número inteiro n, seu fatorial $n!$ é o produto

$$n \times (n - 1) \times (n - 2) \times \ldots \times 3 \times 2 \times 1$$

Este é o número de modos diferentes de arranjar n objetos em sequência. Por exemplo, o alfabeto inglês com 26 letras pode ser arranjado em

$$26! = 403.291.461.126.605.635.584.000.000$$

sequências diferentes. Em artigos escritos em 1876 e 1885, Henri Brocard

notou que

$4! + 1 = 24 + 1 = 25 = 5^2$
$5! + 1 = 120 + 1 = 121 = 11^2$
$7! + 1 = 5.040 + 1 = 5.041 = 71^2$

ou seja, todos quadrados perfeitos. Ele não encontrou nenhum outro fatorial que virasse um quadrado perfeito quando acrescido de 1, e perguntou se existiriam. O gênio indiano autodidata Srinivasa Ramanujan fez-se a mesma pergunta, independentemente, em 1913. Bruce Berndt e William Galway usaram um computador em 2000 para mostrar que não existem soluções adicionais para fatoriais até 1 bilhão.

Números perfeitos ímpares

Um número é perfeito se for igual à soma de seus divisores (isto é, números que o dividem exatamente, excluindo o próprio número). Eis exemplos:

$6 = 1 + 2 + 3$
$28 = 1 + 2 + 4 + 7 + 14$

Euclides provou que se $2^n - 1$ é primo, então $2^{n-1}(2^n - 1)$ é perfeito. Os exemplos acima correspondem a $n = 2$ e 3. Números primos desse formato são chamados primos de Mersenne, e são conhecidos 47. Até hoje, o maior deles é $2^{43.112.609} - 1$, sendo que também é o maior primo conhecido.[1] Euler provou que todos os números perfeitos pares devem ser dessa forma, mas ninguém jamais encontrou um número perfeito ímpar, nem provou que não podem existir. Pomerance deduziu um argumento não rigoroso que indica que não existem. Um número perfeito ímpar deve satisfazer diversas condições rígidas. Deve ser, no mínimo, 10^{300}, deve ter um fator primo maior que 10^8, seu segundo maior fator primo deve ser, no mínimo, 10^4 e precisa ter pelo menos 75 fatores primos e pelo menos doze fatores primos distintos.

Conjectura de Collatz

Pegue um número inteiro. Se for par, divida por 2. Se for ímpar, multiplique por 3 e some 1. Repita indefinidamente. O que acontece?

Por exemplo, comecemos por 12. Números sucessivos são

$$12 \to 6 \to 3 \to 10 \to 5 \to 16 \to 8 \to 4 \to 2 \to 1$$

após o que a sequência $4 \to 2 \to 1 \to 4 \to 2 \to 1$ se repete para sempre. A conjectura de Collatz afirma que o mesmo resultado final ocorre qualquer que seja o número com o qual comecemos. O nome foi dado em homenagem a Lothar Collatz, que a apresentou em 1937, mas ela possui muitos outros nomes: conjectura $3n + 1$, problema do granizo, conjectura de Ulam, problema de Kakutani, conjectura de Thwaites, algoritmo de Hasse e problema de Siracusa.

O que torna o problema difícil é que os números podem muitas vezes explodir. Por exemplo, se começamos com 27, então a sequência sobe até 9.232; mesmo assim, ela finalmente cai para 1 após 111 passos. Simulações de computador verificaram a conjectura para todos os números iniciais até 5.764×10^{18}. Foi provado que nenhum outro ciclo a não ser $4 \to 2 \to 1$ existe que envolva menos do que 35.400 números. A possibilidade de algum número inicial levar a uma sequência que contenha números cada vez maiores, separados por números menores, não foi descartada. Ilia Krasikov e Jeffrey Lagarias provaram que para valores iniciais até n, pelo menos uma constante vezes $n^{0,84}$ deles acabam chegando a 1. Logo, exceções, se existirem, são raras.[2]

Existência de cuboides perfeitos

Esta questão toma como ponto de partida a existência de, e a fórmula para, trincas pitagóricas, e passa o problema para a terceira dimensão. Um tijolo de Euler é um cuboide – um bloco com forma de tijolo – com arestas inteiras, em que todas as diagonais das faces são inteiras. O menor tijolo de Euler foi descoberto em 1719 por Paul Halcke. Suas arestas são 240, 117 e 4; as diagonais das faces valem 267, 244 e 125. Euler descobriu fórmulas para esses tijolos, análogas às fórmulas para trincas pitagóricas, mas estas não dão todas as soluções.

Não se sabe se existe um cuboide perfeito: isto é, se há algum tijolo de

Euler cuja diagonal principal, que atravessa o interior do tijolo de um vértice até o vértice totalmente oposto, também tem medida inteira. (Há quatro dessas diagonais, mas todas têm o mesmo comprimento.) Sabe-se que a fórmula de Euler não consegue fornecer um exemplo. Esse tijolo, se existir, precisa satisfazer diversas condições – por exemplo, pelo menos uma das arestas precisa ser múltipla de 5, uma precisa ser múltipla de 7, uma precisa ser múltipla de 11, e uma precisa ser múltipla de 19. Pesquisas no computador mostraram que um dos lados precisa estar na casa de pelo menos 1 trilhão.

Há alguns resultados que erram por pouco. O tijolo com lados 672, 153 e 104 tem uma diagonal principal inteira e duas das três diagonais das faces também são inteiras. Em 2004, Jorge Sawyer e Clifford Reiter provaram que o paralelepípedo perfeito existe.[3] Um paralelepípedo é como um cuboide, mas suas faces podem ser paralelogramos oblíquos, ou seja, é um cuboide inclinado. As arestas têm comprimentos 271, 106 e 103; as menores diagonais de face têm comprimentos 101, 266 e 255; as maiores diagonais de face têm comprimentos 183, 312 e 323; e as diagonais do corpo têm comprimentos 374, 300, 278 e 272.

Conjectura do corredor solitário

Esta provém de uma obscura área da matemática conhecida como teoria de aproximação diofantina, e foi formulada por Jörg Wills em 1967. Luis Goddyn cunhou o nome em 1998. Suponha que n corredores estejam correndo numa pista circular de comprimento unitário com velocidade constante, sendo que as velocidades são todas diferentes umas das outras. Será que em algum instante todo corredor estará solitário – ou seja, estará a uma distância $1/n$ de todos os outros? Instantes diferentes para corredores diferentes, é claro. A conjectura é que a resposta é sempre "sim", e foi provada quando $n = 4, 5, 6$ e 7.

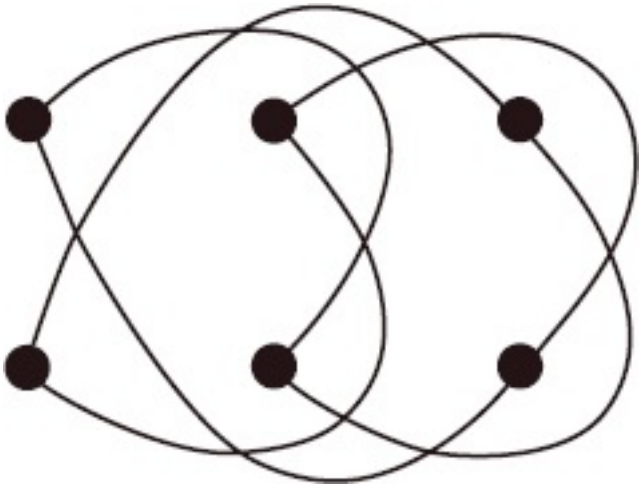


FIGURA 48 Exemplo de *thrackle*.

Conjectura do *thrackle* de Conway

Um *thrackle* é um grafo desenhado no plano de maneira que duas arestas se encontrem exatamente uma vez (Figura 48). Elas podem se encontrar num vértice ou podem se cruzar em pontos internos, mas não ambas as coisas. Caso se cruzem, devem fazê-lo transversalmente, isto é, nenhuma delas deve permanecer do mesmo lado da outra (o que poderia acontecer, por exemplo, caso se tangenciassem mutuamente). Em um trabalho não publicado, John Horton Conway conjecturou que em qualquer *thrackle* o número de arestas é menor ou igual ao número de vértices. Em 2011, Radoslav Fulek e János Pach provaram que todo *thrackle* com n vértices tem, no máximo, $1{,}428n$ arestas.[4]

Irracionalidade da constante de Euler

Não se conhece "forma fechada" alguma para a soma da série harmônica

$$H_n = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \ldots + \frac{1}{n}$$

e provavelmente tal fórmula não exista. No entanto, há uma aproximação excelente: quando n aumenta, H_n vai ficando mais e mais próximo de $\ln n + \gamma$. Aqui γ é a constante de Euler, com valor numérico de aproximadamente

0,5772156649. Euler estabeleceu sua fórmula em 1734, e Lorenzo Mascheroni estudou a constante em 1790. Nenhum deles usou o símbolo γ.

A constante de Euler é um desses números estranhos que às vezes surgem em matemática, como π e *e*, que surgem por toda parte, mas parecem ser criaturas com vida própria, não exprimíveis de nenhuma maneira simpática em termos de números mais simples. Vimos no Capítulo 3 que tanto π como *e* são transcendentais: não resolvem equação algébrica alguma com coeficientes inteiros. Em particular, são irracionais: não são frações exatas. Acredita-se amplamente que a constante de Euler seja transcendental, mas nem sequer sabemos com certeza se é irracional. Se $\gamma = {}^{p}/_{q}$ para *p* e *q* inteiros, então *q* é, no mínimo, $10^{242.080}$.

A constante de Euler é importante em muitas áreas da matemática, desde a função zeta de Riemann até a teoria quântica de campo. Ela surge em muitos contextos e aparece em muitas fórmulas. É ultrajante que não consigamos decidir se ela é racional.

Campos de números quadráticos reais

No Capítulo 7 vimos que alguns campos numéricos algébricos têm decomposição única em fatores primos e outros não. Os campos numéricos algébricos mais bem entendidos são os quadráticos, obtidos quando se pega a raiz quadrada de algum número *d* que não é quadrado perfeito; de fato, ele não tem fatores primos quadrados. O correspondente anel de inteiros algébricos consiste de todos os números da forma $a + b\sqrt{d}$, onde *a* e *b* são inteiros se *d* não for da forma $4k + 1$, e são ou inteiros ou ambos inteiros ímpares divididos por 2, se *d* não for dessa forma.

Quando *d* é negativo, sabe-se que a decomposição em fatores primos é única para exatamente nove valores: −1, −2, −3, −7, −11, −19, −43, −67 e −163. Provar a fatoração única nesses casos é relativamente simples, mas descobrir se há outros é bem mais difícil. Em 1934, Hans Heilbronn e Edward Linfoot mostraram que, no máximo, mais um inteiro negativo pode ser acrescentado à lista. Kurt Heegner deu em 1952 uma prova de que a lista está completa, mas julgou-se que ela possuía uma lacuna. Em 1967, Harold Stark deu uma prova completa, observando que não diferia significativamente da prova de Heegner – ou seja, a lacuna não tinha

importância. Na mesma época, Alan Baker achou uma prova diferente.

O caso quando *d* é positivo é bem distinto. A fatoração é única para muito mais valores de *d*. Até 50, esses valores são 2, 3, 5, 6, 7, 11, 13, 14, 17, 19, 21, 22, 23, 29, 31, 33, 37, 38, 41, 43, 46, 47, e os cálculos de computador revelam muitos outros. Pelo que sabemos, pode haver infinitos *d* positivos para os quais o correspondente campo numérico quadrático tenha fatoração única. Uma análise heurística de Cohen e Lenstra sugere que aproximadamente três quartos de todos os *d* positivos devam definir campos numéricos com fatoração única. Os resultados de computador concordam com essa estimativa. O problema é provar que essas observações estão corretas.

A formiga de Langton

À medida que se desenrola o século XXI, tem ficado cada vez mais visível que algumas das técnicas tradicionais de modelagem matemática são incapazes de lidar com as complexidades dos problemas com que a humanidade se depara, tais como o sistema financeiro global, a dinâmica de ecossistemas e o papel dos genes no crescimento de organismos vivos. Muitos desses sistemas envolvem grandes números de agentes – pessoas, empresas, organismos, genes – que interagem entre si. Essas interações podem muitas vezes ser modeladas com bastante precisão usando-se regras simples. No decorrer dos últimos trinta anos surgiu um novo tipo de modelo, que tenta atacar de frente o comportamento de sistemas com muitos agentes. Para compreender como 100 mil pessoas vão se movimentar por um estádio de esportes, por exemplo, você não tira a média delas para criar uma espécie de fluido humano e indagar como ele flui. Em vez disso, cria um modelo computadorizado com 100 mil agentes individuais, impõe regras adequadas e roda uma simulação para ver o que faz essa multidão de computador. Esse tipo de modelo é chamado sistema complexo.

Para dar-lhe um vislumbre dessa fascinante área nova da matemática, vou descrever um dos sistemas complexos mais simples e explicar por que não o entendemos plenamente. Chama-se formiga de Langton. Christopher Langton foi um dos primeiros membros do Instituto Santa Fé, fundado em 1984 por cientistas como George Cowan e Murray Gell-Mann, entre outros,

para promover a teoria e aplicação de sistemas complexos. Langton inventou sua formiga em 1986. Tecnicamente é um autômato celular, um sistema de células numa grade quadrada, cujos estados são mostrados por cores. A cada passo, a cor da célula muda de acordo com as cores de suas vizinhas.

As regras são absurdamente simples. A formiga vive numa grade quadrada infinita de células, a princípio todas brancas. Ela carrega um pote inesgotável de tinta preta de secagem rápida e outro pote inesgotável de tinta branca de secagem rápida. Ela pode estar virada para o norte, leste, sul ou oeste; por simetria vamos admitir que ela comece virada para o norte. A cada instante ela olha a cor do quadrado que ocupa e a muda de branco para preto ou de preto para branco, usando seus potes de tinta. Se o quadrado era branco, ela então se vira noventa graus para a direita e dá um passo adiante. Se o quadrado era preto, ela então se vira noventa graus para a esquerda e dá um passo adiante. E repete esse comportamento indefinidamente.

Se você simula a formiga,[5] ela começa pintando composições simples, bastante simétricas de quadrados pretos e brancos. De tempos em tempos ela retorna a um quadrado que já visitou, mas sua viagem não se torna repetida porque a cor desse quadrado mudou, então ela se vira para o outro lado na repetição da visita. À medida que a simulação prossegue, a composição da formiga torna-se caótica e aleatória. Ela não tem padrão discernível; basicamente é uma grande bagunça. Nesse estágio você pode imaginar, de forma razoável, que esse comportamento caótico vá continuar indefinidamente. Afinal, quando a formiga revisita uma região caótica ela fará uma série caótica de viradas e repinturas. Se você seguisse adiante com a simulação, os 10 mil passos seguintes, ou algo assim, pareceriam justificar essa conclusão. No entanto, se você continuar, aparece um padrão. A formiga entra num ciclo repetitivo de 104 passos, no fim do qual ela se moveu diagonalmente dois quadrados. Então, ela pinta uma larga faixa diagonal de células pretas e brancas chamada estrada, que continua para sempre (Figura 49).

FIGURA 49 Estrada da formiga de Langton.

Tudo que foi descrito até agora pode ser provado com todo o rigor simplesmente listando-se os passos que a formiga dá. A prova seria bastante longa – uma lista de 10 mil passos –, mas ainda assim seria uma prova. Mas a matemática fica mais interessante se nos fizermos uma pergunta um pouco mais genérica. Suponha que antes de a formiga partir pintemos de preto um número finito de células. Podemos escolher esses quadrados como bem quisermos: pontos ao acaso, um retângulo sólido, a Mona Lisa. Podemos pintar 1 milhão deles, ou 1 bilhão, mas não infinitos. O que acontece?

As excursões iniciais da formiga mudam drasticamente toda vez que ela dá de encontro com um dos nossos novos quadrados pretos. Ela pode perambular por todos os lados, desenhando e redesenhando formas intrincadas … Mas em toda simulação até hoje realizada, não importa qual tenha sido a configuração inicial, a formiga acaba construindo sua estrada, usando o mesmo ciclo de 104 passos. E isso sempre acontece? Essa estrada é o "atrator" especial da dinâmica da formiga? Ninguém sabe. É um dos problemas básicos não solucionados da teoria da complexidade. O máximo que sabemos é que qualquer que seja a configuração inicial de células pretas, a formiga permanece sempre dentro de uma região delimitada da grade.

Conjectura da matriz de Hadamard

Uma matriz de Hadamard, que recebeu o nome de Jacques Hadamard, é um arranjo quadrado de 0s e 1s de tal maneira que quaisquer duas linhas ou colunas distintas concordem em metade de suas disposições e discordem na outra metade. Usando preto e branco para indicar 1 e 0, a Figura 50 mostra matrizes de Hadamard de tamanhos 2, 4, 8, 12, 16, 20, 24 e 28. Essas matrizes aparecem em muitos problemas matemáticos e em ciência da computação, especialmente em teoria de codificação. (Em algumas aplicações, entre elas a motivação original de Hadamard, os quadrados brancos correspondem a −1, e não a 0.)

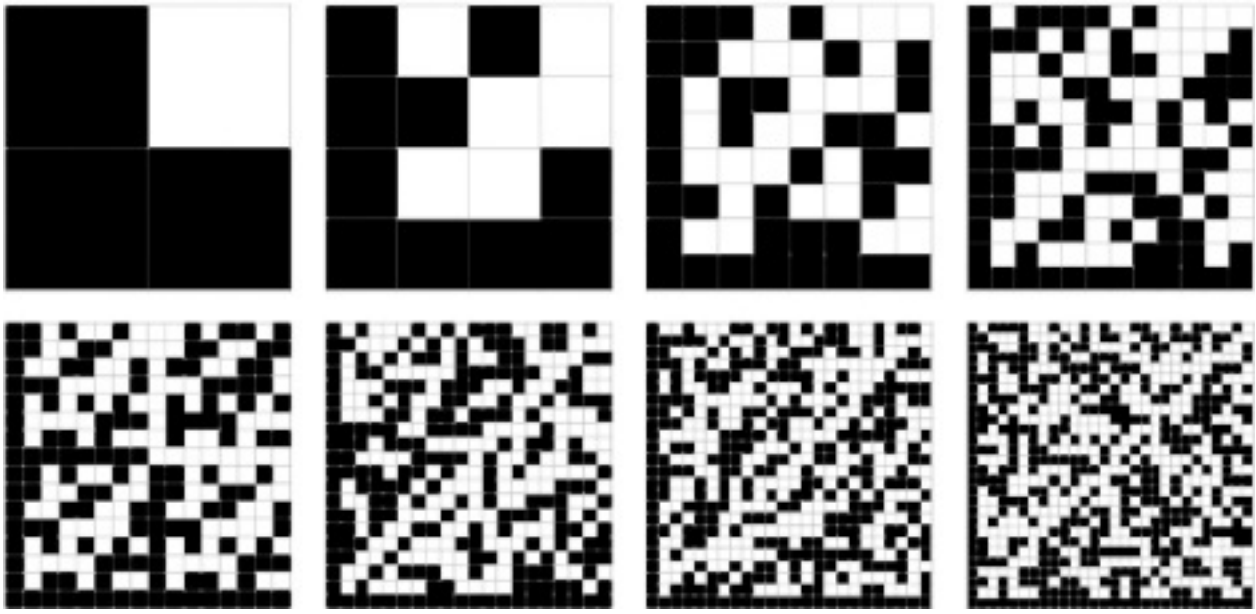


FIGURA 50 Matrizes de Hadamard tamanhos 2, 4, 8, 12, 16, 20, 24 e 28.
http://mathworld.wolfram.com/HadamardMatrix.html

Hadamard provou que tais matrizes existem apenas quando $n = 2$ ou n é múltiplo de 4. O teorema de Paley, de 1933, prova que uma matriz de Hadamard sempre existe se o tamanho da matriz é múltiplo de 4 e igual a $2^a(p^b + 1)$, onde p é um primo ímpar. Múltiplos de 4 não cobertos por este teorema são 92, 116, 156, 172, 184, 188, 232, 236, 260, 268 e outros valores maiores. A conjectura afirma que uma matriz de Hadamard existe sempre que o tamanho for um múltiplo de 4. Em 1985, K. Sawade encontrou uma de tamanho 268; os outros números não cobertos pelo teorema de Paley já haviam sido abordados. Em 2004, Hadi Kharaghani e Behruz Tayfeh-Rezaie acharam uma matriz de Hadamard de tamanho 428, e o menor tamanho para o qual não se conhece solução é 668.

Equação de Fermat-Catalan

Esta é a equação diofantina $x^a + y^b = z^c$ onde a, b e c são inteiros positivos, os expoentes. Eu a chamarei equação de Fermat-Catalan porque sua solução relaciona-se tanto com o último teorema de Fermat (Capítulo 7) como com a conjectura de Catalan (Capítulo 6). Se a, b e c são pequenos, soluções inteiras não nulas não constituem especial surpresa. Por exemplo, se os três números forem 2, temos uma equação pitagórica, conhecida desde os tempos de Euclides como tendo infinitas soluções. Logo, o principal interesse está nos casos em que esses expoentes são grandes. A definição técnica de "grande" é que $s = {}^1/_a + {}^1/_b + {}^1/_c$ seja menor que 1. Apenas dez soluções grandes da equação de Fermat-Catalan são conhecidas:

$1 + 2^3 = 3^2$

$2^5 + 7^2 = 3^4$

$7^3 + 13^2 = 2^9$

$2^7 + 17^3 = 71^2$

$3^5 + 11^4 = 122^2$

$17^7 + 76271^3 = 21063928^2$

$1414^3 + 2213459^2 = 65^7$

$9262^3 + 15312283^2 = 113^7$

$43^8 + 9622^3 = 30042907^2$

$33^8 + 159034^2 = 15613^3$

O primeiro deles é considerado grande porque $1 = 1^a$ para qualquer a, e $a = 7$ satisfaz a definição. A conjectura de Fermat-Catalan afirma que a equação de Fermat-Catalan tem apenas um número inteiro finito de soluções, sem um fator comum, quando s é grande. O principal resultado foi provado em 1997 por Henri Darmon e Loïc Merel: não há soluções nas quais $c = 3$ e a e b sejam iguais ou maiores que 3. Pouco mais se sabe. Progresso adicional parece depender de uma nova conjectura fascinante, que vem a seguir.

Conjectura ABC

Em 1983, Richard Mason notou que um caso do último teorema de Fermat fora ignorado: potências de primeiro grau. Ou seja, considerar a equação $a + b = c$.

À primeira vista a ideia é absolutamente sem sentido. É preciso muito pouca compreensão de álgebra para resolver esta equação para qualquer

uma das três variáveis em termos das outras duas. Por exemplo, $a = c - b$. O que muda todo o jogo, porém, é o contexto. Mason percebeu que tudo fica mais profundo se fizermos as perguntas certas a respeito de a, b e c. O resultado dessa ideia extraordinária foi uma nova conjectura em teoria dos números com consequência de longo alcance. Seria possível dispor de muitos problemas atualmente não resolvidos e chegar a provas melhores e mais simples de alguns dos maiores teoremas em teoria dos números. Trata-se da conjectura ABC, que é respaldada por uma enorme quantidade de evidências numéricas. Ela repousa sobre uma analogia livre entre inteiros e polinômios.

Euclides e Diofanto conheciam uma receita para trincas pitagóricas, que agora escrevemos como uma fórmula (Capítulo 6). Será que esse artifício pode ser repetido com outras equações? Em 1851, Joseph Liouville provou que não existe tal fórmula para a equação de Fermat quando a potência é 3 ou mais. Mason aplicou raciocínio semelhante à equação mais simples

$$a(x) + b(x) = c(x)$$

para três polinômios. É uma ideia ultrajante, porque todas as soluções podem ser encontradas usando álgebra elementar. O principal resultado, porém, é elegante e longe de ser óbvio: se cada polinômio tem um fator que seja um quadrado, um cubo, ou potência superior, a equação não tem soluções.

Teoremas sobre polinômios frequentemente têm análogos sobre inteiros. Em particular, polinômios irredutíveis correspondem a números primos. O análogo natural para inteiros do teorema de Mason sobre polinômios é o seguinte. Suponha que $a + b = c$, onde a, b e c sejam inteiros sem fator comum; então a quantidade de fatores primos de cada um deles – a, b e c – é menor do que o número de fatores primos *distintos* de abc. Infelizmente, exemplos simples mostram que isso é falso. Em 1985, David Masser e Joseph Oesterlé modificaram a afirmação e propuseram uma versão dessa conjectura que não conflitasse com quaisquer exemplos conhecidos. Sua conjectura ABC pode muito bem ser a maior questão atualmente em aberto em teoria dos números.[6] Se alguém provasse amanhã a conjectura ABC, muitos teoremas profundos e difíceis teriam provas novas e simples. Outra consequência seria a conjectura de Marshall Hall: a diferença entre qualquer cubo perfeito e qualquer quadrado perfeito tem de ser razoavelmente grande. E ainda outra aplicação potencial da conjectura ABC é para o problema de Brocard, o primeiro deste capítulo.

Em 1993, Marius Overholt provou que se a conjectura ABC for verdadeira, existe apenas uma quantidade finita de soluções para a equação de Brocard.

Uma das consequências mais interessantes da conjectura ABC está relacionada com a conjectura de Mordell. Faltings a provou usando métodos sofisticados, mas seu resultado seria ainda mais poderoso se tivéssemos uma peça adicional de informação: um limite quanto ao tamanho das soluções. Aí existiria um algoritmo para encontrá-las todas. Em 1991, Noam Elkies mostrou que uma versão específica da conjectura ABC, na qual várias constantes que aparecem são limitadas, implica este aperfeiçoamento no teorema de Faltings. Laurent Moret-Bailly mostrou, de modo bastante consistente, que o inverso é verdadeiro. Limites suficientemente fortes no tamanho das soluções de *apenas uma* equação diofantina, $y^2 = x^5 - x$, implicam a totalidade da conjectura ABC. Embora não seja tão bem conhecida como muitas outras conjecturas não resolvidas, a ABC é, sem dúvida, um dos grandes problemas da matemática. Segundo Granville e Thomas Tucker, dispor dela teria "um extraordinário impacto sobre a nossa compreensão da teoria dos números. Prová-la ou refutá-la seria assombroso".[7]

Glossário

3-esfera: Análoga tridimensional à esfera: conjunto de todos os pontos num espaço quadridimensional a uma determinada distância de um ponto fixo, o centro.

Algoritmo: Procedimento especificado para solucionar um problema, com garantia de parar ao chegar a uma resposta.

Análise complexa: A análise – cálculo com rigor lógico – executada com funções com valores complexos de uma variável complexa.

Aritmética modular: Um sistema aritmético no qual múltiplos de algum número específico, chamado *módulo*, são tratados como se fossem todos iguais a zero.

Assintóticas: Duas grandezas definidas em termos de uma variável são assintóticas se sua razão se aproxima mais e mais de 1 à medida que a variável se torne tão grande quanto se queira.

Autovalor: Um número num conjunto de números associados com um operador. Se o operador aplicado a um vetor produz um múltiplo constante desse vetor, o referido múltiplo é o seu autovalor.

Bola: Uma esfera sólida – ou seja, a esfera e seu interior.

Bóson de Higgs: Partícula fundamental cuja existência explica por que todas as partículas têm massa. Sua descoberta pelo Grande Colisor de Hádrons foi anunciada em julho de 2012.

Campo de velocidade: Função que especifica a velocidade em cada ponto do espaço. Por exemplo, quando um fluido flui, sua velocidade pode ser especificada em cada ponto, assumindo valores diferentes em pontos diferentes.

Campo eletromagnético: Função que especifica as intensidades e direções dos campos elétrico e magnético em qualquer ponto do espaço.

Caos: Comportamento aparentemente aleatório num sistema determinístico.

Característica de Euler: $F - A + V$, onde F é o número de faces na triangulação de algum espaço, A é o número de arestas e V é o número de vértices. Para um toro com g furos, é igual a $2 - 2g$, qualquer que seja a triangulação.

Ciclo: Em topologia: uma combinação formal de laçadas numa triangulação com rótulos numéricos anexados. Em geometria algébrica: uma combinação formal de subvariedades com rótulos numéricos anexados.

Classe de Hodge: Uma classe de co-homologia de ciclos numa variedade algébrica com propriedades analíticas especiais.

Classe E: Algoritmo cujo tempo de processamento, para um *input* de tamanho n, seja comparável à enésima potência de alguma constante.

Classe não-P: Não classe P.

Classe NP: Problema para o qual uma solução proposta pode ser verificada (mas não necessariamente encontrada) por um algoritmo classe P.

Classe P: Algoritmo cujo tempo de processamento seja comparável a alguma potência fixa do tamanho do *input*.

Coeficiente: Em um polinômio como $6x^3 - 5x^2 + 4x - 7$, os coeficientes são os números 6, −5, 4, −7, que multiplicam as várias potências de x.

Configuração inevitável: Membro de uma lista de subgrafos, pelo menos um dos quais deve ocorrer em qualquer grafo no plano.

Configuração redutível: Parte de um grafo com a seguinte propriedade: se o grafo obtido

removendo-o pode ser colorido com quatro cores, então o grafo original também pode.

Conjunto: Coleção de objetos (matemáticos). Por exemplo, o conjunto de todos os números inteiros.

Constante de Euler: Um número especial representado por γ, aproximadamente igual a 0,57721. Ver nota 7, Capítulo 9.

Construção com regra e compasso: Qualquer construção geométrica que possa ser executada usando-se uma régua não graduada e um compasso (falando corretamente: um par de compassos).

Contraexemplo: Um exemplo que refuta uma afirmação. Podemos citar: 9 é um contraexemplo de "todos os números ímpares são primos".

Contraexemplo mínimo: Um objeto matemático que não possui alguma propriedade desejada, e em certo sentido é o menor possível entre esses objetos. Por exemplo, um mapa que não pode ser colorido com quatro cores, e também o menor número de regiões para as quais isso pode ocorrer. Contraexemplos mínimos são frequentemente hipotéticos, e o objetivo é provar que não existem.

Coordenada: Um número numa lista que determina a posição de um ponto em um plano ou no espaço.

Cosseno: Uma função trigonométrica de um ângulo, definida por $\cos A = {}^{a}/_{c}$ na Figura 51.

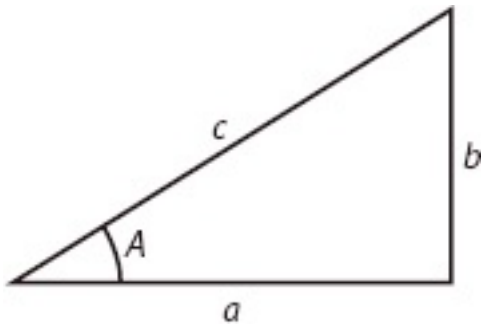


FIGURA 51 Cosseno (${}^{a}/_{c}$), seno (${}^{b}/_{c}$) e tangente (${}^{b}/_{a}$) de um ângulo A.

Cubo: Um número multiplicado por si mesmo e novamente por si mesmo. Por exemplo, o cubo de 7 é $7 \times 7 \times 7 = 343$. Geralmente escrito como 7^3.

Curva elíptica: Curva no plano cuja equação tem a forma $y^2 = ax^3 + bx^2 + cx + d$ para constantes a, b, c, d, normalmente assumindo ser racional. Ver Figura 27.

Curvatura: Medida de quanto o espaço se curva próximo a um ponto dado. Uma esfera tem curvatura positiva, um plano tem curvatura zero, e um espaço em forma de sela tem curvatura negativa.

Dimensão: O número de coordenadas necessárias para especificar a localização de um ponto num determinado espaço. Por exemplo, o plano tem dimensão 2 e o espaço em que vivemos (conforme modelado pela geometria de Euclides) tem dimensão 3.

Disco (topológico): Região em uma superfície que pode ser deformada continuamente num círculo mais seu interior.

Dodecaedro: Sólido cujas faces são doze pentágonos regulares. Ver Figura 38.

Eixo de rotação: Uma reta fixa em torno da qual o objeto gira.

Empilhamento: Coleção de formas arranjadas no espaço de modo que não se sobreponham.

Empilhamento reticulado: Coleção de círculos ou esferas idênticos cujos centros formam um reticulado.

Equação cúbica (ou de terceiro grau): Qualquer equação $ax^3 + bx^2 + cx + d = 0$, onde x é desconhecido (a incógnita) e a, b, c, d são constantes.

Equação diferencial: Equação que relaciona uma função com a sua taxa de variação.

Equação diferencial parcial: Equação diferencial envolvendo as taxas de variação de alguma função em relação a duas ou mais variáveis (com frequência espaço e tempo).

Equação diofantina: Equação para a qual se requer que as soluções sejam números racionais.

Equação quadrática (ou do segundo grau): Qualquer equação $ax^2 + bx + c = 0$, onde x é desconhecido (a incógnita) e a, b, c são constantes.

Esfera: Conjunto de todos os pontos no espaço a uma determinada distância de um ponto fixo, o centro. Ela é redonda, como uma bola, mas o termo "esfera" refere-se somente aos pontos na superfície da bola, e não no seu interior.

Espaço topológico: Forma que é considerada como "a mesma" se for sujeita a qualquer transformação contínua.

Estável: Estado de um sistema dinâmico ao qual ele retorna se sujeito a uma pequena perturbação.

Expoente: Numa potência da variável x, o expoente é a referida potência. Por exemplo, em x^7 o expoente é 7.

Fase: Número complexo no círculo unitário usado para multiplicar uma função de onda quântica.

Fatoração única em primos: Propriedade de qualquer número poder ser escrito como produto de números primos de uma única maneira, exceto mudando-se a ordem em que os fatores são escritos. Esta propriedade é válida para inteiros, mas pode falhar em sistemas algébricos mais gerais.

Fatoração, ou decomposição em fatores primos: O processo que escreve um número em termos de seus divisores primos. Por exemplo, a fatoração de 60 em primos é $2^2 \times 3 \times 5$.

Fluxo de Ricci: Equação que prescreve como a curvatura do espaço varia com o tempo.

Fronteira: As bordas de uma região específica.

Função: Uma regra f que, quando aplicada a um número x, produz outro número $f(x)$. Por exemplo, se $f(x) = \ln x$, então f é a função logarítmica. A variável x pode ser real ou complexa (nesse caso costuma ser escrita como z). De forma mais geral, x e $f(x)$ podem ser membros de conjuntos específicos; em particular, o plano ou o espaço.

Função de onda quântica: Função matemática determinando as propriedades de um sistema quântico.

Função elíptica: Função complexa que permanece inalterada quando dois números complexos independentes são somados à sua variável. Isto é, $f(z) = f(z + u) = f(z + v)$, onde v não é um múltiplo real de u. Ver Figura 30.

Função-L de Dirichlet: Uma generalização da função zeta de Riemann.

Função zeta: Função complexa introduzida por Riemann que representa analiticamente os números primos. É definida pela série

$$\zeta(s) = \frac{1}{1^s} + \frac{1}{2^s} + \frac{1}{3^s} + \frac{1}{4^s} + \frac{1}{5^s} + \frac{1}{6^s} + \frac{1}{7^s} + \dots$$

que converge quando a parte real de s é maior que 1. Esta definição pode ser estendida a todo s complexo, exceto 1, por um processo chamado continuação analítica.

Genus: Número de furos numa superfície.

Geometria não euclidiana: Alternativa para a geometria de Euclides na qual as propriedades habituais de pontos e retas permanecem válidas, exceto pela existência da paralela única a uma reta dada passando por um ponto dado. Há dois tipos: elíptica e hiperbólica.

Geometria projetiva: Espécie de geometria na qual retas paralelas não existem: quaisquer duas retas encontram-se num único ponto. É obtida por meio da geometria euclidiana adicionando-se uma nova "linha do infinito".

Grafo: Conjunto de pontos (nós) ligados por traços (arestas).

Grafo dual: Um grafo obtido a partir de um grafo dado, associando-se a cada região um vértice, e

ligando os vértices por arestas se as regiões correspondentes forem adjacentes. Ver Figura 10.

Grau: A maior potência de uma variável que ocorre num polinômio. Por exemplo, o grau de $6x^3 - 5x^2 + 4x - 7$ é 3.

Grupo de co-homologia: Estrutura algébrica abstrata associada com um espaço topológico, análogo mas "dual" ao grupo de homologia.

Grupo fundamental: O grupo formado por classes de homotopia para laçadas em algum espaço topológico, sob a operação "percorra a primeira laçada e depois a segunda".

Grupo trivial: Grupo que consiste apenas de um único elemento, a identidade.

Grupo: Estrutura algébrica abstrata compreendendo um conjunto e uma regra para combinar dois elementos quaisquer do conjunto, sujeita a três condições: a propriedade associativa, a existência de um elemento de identidade e a existência de inversos.

Homologia (grupo de): Invariante topológico de um espaço, definido por laçadas fechadas. Duas dessas laçadas são homólogas se sua diferença for a fronteira de um disco topológico.

Homotopia (grupo de): Invariante topológico de um espaço, definido por laçadas fechadas. Duas dessas laçadas são homotópicas se cada uma puder ser deformada continuamente na outra.

Ideal (número): Número que não está contido num dado sistema de números algébricos, mas está relacionado com esse sistema de uma forma que restaure a fatoração única em primos nos casos em que essa propriedade falha. Substituído em álgebra moderna por um ideal, que é um tipo especial de subconjunto do referido sistema.

Ideal primo: Análogo de um número primo para sistema de números algébricos.

Indução: Método geral para provar teoremas sobre números inteiros. Se alguma propriedade é válida para 0, e sua validade para qualquer número inteiro n implique a validade para $n + 1$, então a propriedade é válida para todos os números inteiros.

Instável: Estado de um sistema dinâmico ao qual ele pode não retornar se for sujeito a uma pequena perturbação.

Integral: Operação do cálculo que efetivamente soma grandes quantidades de pequenas contribuições. A integral de uma função é a área sob o gráfico.

Integral logarítmica: A função $\mathrm{Li}(x) = \int_0^x \frac{\mathrm{d}t}{\ln t}$

Inteiro: Qualquer um dos números ..., −3, −2, −1, 0, 1, 2, 3, ...

Inteiro algébrico: Número complexo que satisfaz uma equação polinomial com coeficientes inteiros e o coeficiente do termo da maior potência igual a 1. Por exemplo, $i\sqrt{2}$, que satisfaz a equação $x^2 + 2 = 0$.

Inteiro ciclotômico/número: Soma de potências de uma raiz complexa da unidade com coeficientes inteiros/racionais.

Laçada: Uma curva fechada num espaço topológico.

Limite superior: Número específico que é garantidamente maior que qualquer grandeza cujo tamanho está sendo pesquisado.

Logaritmo: O logaritmo (natural) de x, escrito ln x, é a potência à qual e (= 2,71828...) deve ser elevado para obter x. Isto é, $e^{\ln x} = x$.

Máximo: O maior valor de algo.

Mínimo: O menor valor de algo.

Modelo Padrão: Modelo em mecânica quântica que explica todas as partículas fundamentais conhecidas.

Momento angular: Medida de quanta rotação um corpo possui.

NP-completa: Uma classe específica de problema NP, com a propriedade de que se existir um algoritmo classe P para resolvê-lo, então *qualquer* problema NP pode ser solucionado usando um algoritmo classe P.

Número algébrico: Número complexo que satisfaz uma equação polinomial com coeficientes inteiros, ou coeficientes racionais equivalentes. Por exemplo, $i\sqrt{2/3}$, que satisfaz a equação $x^2 + {}^2/_9$

= 0, ou, equivalentemente, $9x^2 + 2 = 0$.

Número complexo: Número da forma $a + bi$, onde i é a raiz quadrada de menos um e a, b são números reais.

Número composto: Número inteiro que pode ser obtido multiplicando-se dois números inteiros menores entre si.

Número congruente: Número que pode ser a diferença comum de uma sequência dos quadrados de três números racionais.

Número de Fermat: Número da forma $2^{2^k}+1$, onde k é um número inteiro. Se este número é primo então é chamado de *primo de Fermat*.

Número de voltas: Número de vezes que uma curva dá volta no sentido anti-horário em torno de algum ponto escolhido.

Número irracional: Um número real que não é racional, ou seja, não é da forma p/q onde p e q são inteiros e $q \neq 0$. Exemplos são $\sqrt{2}$ e π.

Número natural: São os inteiros positivos 0, 1, 2, 3, … .

Número primo: Um número inteiro maior que 1 que não pode ser obtido multiplicando-se dois números inteiros menores. Os primeiros números primos são 2, 3, 5, 7, 11, 13.

Número racional: Número real da forma p/q, onde p e q são inteiros e $q \neq 0$. Um exemplo é $22/7$.

Número real: Qualquer número que pode ser expresso em decimais, mesmo podendo continuar para sempre – por exemplo, π = 3,14159265358979323 85 … .

Número transcendental: Número que não satisfaz qualquer equação algébrica com coeficientes racionais. Exemplos são π e e.

Onda: Perturbação que se move através de um meio sólido, líquido ou gasoso sem provocar qualquer mudança permanente no meio.

Operador: Tipo especial de função A, que quando aplicada a um vetor v produz outro vetor Av. Ele deve satisfazer as condições de linearidade $A(v + w) = Av + Aw$ e $A(av) = aA(v)$ para qualquer constante a.

Ordem: O maior número de soluções racionais independentes da equação que define uma curva elíptica. "Independente" significa que elas não podem ser deduzidas de outras soluções usando uma construção geométrica padrão que combina duas soluções quaisquer para gerar uma terceira. Ver Figura 25.

Otimização: Achar o máximo ou mínimo de alguma função.

Partícula: Massa concentrada em um ponto.

Pentágono: Polígono de cinco lados.

Periódico: Qualquer coisa que repete o mesmo comportamento indefinidamente.

Poliedro: Sólido cuja fronteira consiste de um número finito de polígonos.

Polígono: Forma plana cuja fronteira consiste de um número finito de segmentos retos.

Polígono regular: Polígono cujos lados têm todos o mesmo comprimento e cujos ângulos são iguais. Ver Figura 4.

Polinômio: Expressão algébrica como $6x^3 - 5x^2 + 4x - 7$, na qual as potências da variável x são multiplicadas por constantes e somadas entre si.

Polinômio irredutível: Um polinômio que não pode ser obtido multiplicando-se dois polinômios de graus menores.

Potência: Um número multiplicado por si mesmo um número específico de vezes. Por exemplo, a quarta potência de 3 é $3 \times 3 \times 3 \times 3 = 81$, simbolizada como 3^4.

Progressão aritmética: Sequência de números na qual cada número sucessivo é o anterior mais um valor fixo, a diferença comum, chamada *razão* da progressão. Por exemplo, 2, 5, 8, 11, 14, … com a diferença comum de 3.

Quadrado: Um número multiplicado por si mesmo. Por exemplo, o quadrado de 7 é $7 \times 7 = 49$, simbolizado como 7^2.

Quantidade de movimento: Massa multiplicada pela velocidade.

Raiz da unidade: Um número complexo ζ para o qual alguma potência ζk é 1. Ver Figura 7 e nota 2, Capítulo 7.

Razão: A razão entre dois números a e b é $^{a}/_{b}$.

Relatividade geral: Teoria de Einstein da gravitação, na qual a força da gravidade é interpretada como a curvatura do espaço-tempo.

Reticulado: No plano: um conjunto de pontos que repete sua forma ao longo de duas direções independentes, como padrões de papel de parede. Ver Figura 26. No espaço: um conjunto de pontos que repete sua forma ao longo de três direções independentes, como átomos num cristal.

Reticulado cúbico de faces centradas: Um conjunto repetido de pontos no espaço, obtido empilhando cubos uns sobre os outros num tabuleiro tridimensional, e então pegando os cantos dos cubos e os centros de suas seis faces quadradas. Ver Figuras 17 e 19.

Rombidodecaedro: Sólido cuja fronteira é composta de doze losangos idênticos – paralelogramos com os quatro lados iguais. Ver Figura 15.

Rotação: No plano: uma transformação na qual todos os pontos se movem de um mesmo ângulo em torno de um centro fixo. No espaço: uma transformação na qual todos os pontos se movem de um mesmo ângulo em torno de uma reta fixa, o eixo.

Seno: Função trigonométrica de um ângulo, definida por sen $A = {}^{b}/_{c}$ na Figura 51.

Sequência: Lista de números dispostos em ordem. Por exemplo, a sequência 1, 2, 4, 8, 16, … das potências de 2.

Série: Expressão na qual muitas grandezas – frequentemente infinitas – são somadas entre si.

Série de potências: O mesmo que um polinômio exceto que podem ocorrer infinitas potências da variável – por exemplo, $1 + 2x + 3x^2 + 4x^3 + \ldots$ Em circunstâncias adequadas esta soma infinita pode adquirir um valor bem-definido, e a série é dita convergente.

Simetria: Transformação de algum objeto que mantém inalterada sua forma geral. Por exemplo, girar um quadrado em um ângulo reto.

Simetria de calibre: Grupo de simetrias locais de um sistema de equações: transformações das variáveis que podem variar de ponto para ponto no espaço, com a propriedade de que qualquer solução das equações se mantenha solução contanto que seja feita nas equações uma alteração compensatória com interpretação física coerente.

Singularidade: Ponto em que ocorre algo estranho, tal como uma função tornar-se infinita ou a solução de alguma equação deixar de existir.

Sistema dinâmico: Qualquer sistema que mude com o tempo segundo regras específicas. Por exemplo, o movimento dos planetas no sistema solar.

Sólido regular: Sólido cuja fronteira é composta de polígonos regulares idênticos, dispostos da mesma maneira em cada vértice. Euclides provou que existem exatamente cinco sólidos regulares.

Superfície: Forma no espaço obtida juntando-se regiões topologicamente equivalentes ao interior de um círculo. Exemplos são a esfera e o toro.

Tangente: Função trigonométrica de um ângulo, definida por tan $A = {}^{b}/_{a}$ na Figura 51.

Tempo de explosão: Tempo além do qual a solução para uma equação diferencial deixa de existir.

Teoria de calibre: Teoria quântica de campo com um grupo de simetrias de calibre.

Teoria quântica de campo: Teoria em mecânica quântica de uma grandeza que permeia o espaço, podendo ter (e geralmente tem) diferentes valores em diferentes locais.

Topologia: Estudo dos espaços topológicos.

Toro: Superfície como uma rosquinha com um furo. Ver Figura 12.

Toro plano: Toro obtido fazendo coincidir lados opostos de um quadrado cuja geometria natural tem curvatura zero. Ver Figura 12.

Transformação: Outra palavra para "função", geralmente usada quando as variáveis envolvidas são pontos em algum espaço. Por exemplo, "girar em torno do centro através de um ângulo reto" é

uma transformação de um quadrado.

Transformação contínua: Transformação de um espaço com a propriedade de que pontos muito próximos não são puxados para muito longe um do outro.

Translação: Transformação do espaço no qual todos os pontos se deslocam numa mesma distância e numa mesma direção.

Triangulação: Dividir uma superfície em uma rede de triângulos, ou sua análoga multidimensional.

Trinca pitagórica: Três números inteiros a, b, c tais que $a^2 + b^2 = c^2$. Por exemplo, $a = 3$, $b = 4$, $c = 5$. Pelo teorema de Pitágoras, números desse tipo formam os lados de um triângulo retângulo.

Trissecção: Dividir em três partes iguais, sobretudo em relação a ângulos. *Variável:* Grandeza que pode assumir qualquer valor num domínio.

Variedade: Forma no espaço definida por um sistema de equações polinomiais.

Variedade algébrica: Espaço multidimensional definido por um conjunto de equações algébricas.

Velocidade: Taxa segundo a qual a posição varia em relação ao tempo. A velocidade tem tanto um valor quanto direção e sentido.

Vetor: Em mecânica, uma grandeza com tamanho, direção e sentido. Em álgebra e análise, uma generalização dessa ideia.

Vórtice: Um fluido dando voltas e voltas, como um redemoinho. Pode ter qualquer tamanho, inclusive ser muito pequeno.

Zero (de uma função): Se f é uma função, então x é um zero se $f(x) = 0$.

Notas

Epígrafe

1. O original em alemão é: “Wir müssen wissen. Wir werden wissen.” Ocorre num discurso que Hilbert gravou para o rádio. Ver Constance Reid, *Hilbert*, Springer, Berlim, 1970, p.196.

1. Grandes problemas

1. Simon Singh, *Fermat’s Last Theorem*, Fourth Estate, 1997.
2. Gauss, carta a Heinrich Olbers, 21 mar 1816.
3. O título de Wiles era “Modular curves, elliptic forms, and Galois representations” (Curvas modulares, formas elípticas e representações de Galois).
4. Andrew Wiles, “Modular elliptic curves and Fermat’s last theorem” (Curvas elípticas modulares e o último teorema de Fermat), in *Annals of Mathematics*, n.141 (1995), p.443-551.
5. Ian Stewart, *17 equações que mudaram o mundo*, Capítulo 11.
6. Ibid., Capítulo 9.
7. Os problemas de Hilbert e sua condição atual, ligeiramente editado de *Incríveis passatempos matemáticos*, são os seguintes:

    1. **Hipótese do continuum:** Existe algum número cardinal infinito estritamente entre as cardinalidades dos números inteiros e reais? Solucionado por Paul Cohen em 1963 – a resposta depende de quais axiomas são usados para a teoria estabelecida.
    2. **Consistência lógica da aritmética:** Provar que os axiomas padrões da aritmética nunca podem levar a uma contradição. Solucionado por Kurt Gödel em 1931: impossível com os axiomas usuais para a teoria estabelecida.
    3. **Igualdade de volumes de tetraedros:** Se dois tetraedros têm o mesmo volume, é possível sempre cortar um deles em infinitos pedaços poligonais e remontá-los de modo a formar o outro? Solucionado em 1901 por Max Dehn, e a resposta é negativa.
    4. **Linha reta como a menor distância entre dois pontos:** Formular axiomas para a geometria em termos da definição acima de “linha reta” e investigar as implicações. Amplo demais para ter solução definitiva, porém muito trabalho foi feito.
    5. **Grupos de Lie sem assumir diferenciabilidade:** Questão técnica na teoria dos grupos de transformações. Em uma interpretação, solucionado por Andrew Gleason na década de 1950. Em outra, por Hidehiko Yamabe.
    6. **Axiomas para a física:** Desenvolver um sistema rigoroso de axiomas para áreas matemáticas da física, tais como probabilidade e mecânica. Andrei Kolmogorov estabeleceu axiomas para a probabilidade em 1933.
    7. **Números irracionais e transcendentais:** Provar que certos números são irracionais ou transcendentais. Solucionado por Aleksandr Gelfond e Theodor Schneider em 1934.
    8. **Hipótese de Riemann:** Provar que todos os zeros não triviais da função zeta de Riemann encontram-se sobre a linha crítica. Ver Capítulo 9.
    9. **Leis de reciprocidade em campos numéricos:** Generalizar a lei clássica da reciprocidade

quadrática, sobre quadrados em algum módulo, para potências mais elevadas. Parcialmente solucionado.

10. **Determinar quando uma equação diofantina tem soluções:** Achar um algoritmo que, quando alimentado com uma equação polinomial em muitas variáveis, determine se existem quaisquer soluções em números inteiros. Provado impossível por Yuri Matiyasevich em 1970.
11. **Formas quadráticas com números algébricos como coeficientes:** Assuntos técnicos sobre a solução de equações diofantinas em muitas variáveis. Parcialmente solucionado.
12. **Teorema de Kronecker sobre campos abelianos:** Questões técnicas generalizando um teorema de Kronecker. Ainda não resolvido.
13. **Resolver equações de sétimo grau usando funções especiais:** Provar que a equação geral de sétimo grau não pode ser resolvida usando funções de duas variáveis. Uma interpretação de Andrei Kolmogorov e Vladimir Arnold refutou a afirmação.
14. **Finitude de sistemas completos de funções:** Estender um teorema de Hilbert sobre invariantes algébricos a todos os grupos de transformações. Provado falso por Masayoshi Nagata em 1959.
15. **Cálculo enumerativo de Schubert:** Hermann Schubert encontrou um método não rigoroso para contar várias configurações geométricas. Tornar o método rigoroso. Ainda sem solução completa.
16. **Topologia de curvas e superfícies:** Quantos componentes interligados pode ter uma curva algébrica de determinado grau? Quantos ciclos periódicos distintos pode ter uma equação diferencial algébrica de determinado grau? Progresso limitado.
17. **Expressar formas definidas por quadrados:** Se uma função racional sempre assume valores não negativos, deverá ela ser uma soma de quadrados? Solucionado por Emil Artin, D.W. Dubois e Albrecht Pfister. Verdadeiro para números reais, falso em alguns outros sistemas numéricos.
18. **Ladrilhamento do espaço com poliedros:** Questões gerais sobre preencher o espaço com poliedros congruentes. Menciona também a conjectura de Kepler, agora provada. Ver Capítulo 5.
19. **Analiticidade de soluções em cálculo de variações:** O cálculo de variações responde a perguntas como: "Achar a curva mais curta com as seguintes propriedades." Se este problema for definido por funções simpáticas, a solução também deve ser simpática? Provado por Ennio de Giorgi, em 1957, e por John Nash.
20. **Problemas de valor de fronteira:** Compreender as soluções das equações diferenciais da física, dentro de alguma região do espaço, quando propriedades da solução sobre a fronteira dessa região são prescritas. Essencialmente solucionado por inúmeros matemáticos.
21. **Existência de equações diferenciais com monodromia dada:** Um tipo especial de equação diferencial complexa pode ser entendido em termos de seus pontos singulares e grupo de monodromia. Provar que pode ocorrer qualquer combinação desses dados. Resposta sim ou não, dependendo da interpretação.
22. **Uniformização usando funções automórficas:** Questão técnica sobre simplificar equações. Solucionada por Paul Koebe logo após 1900.
23. **Desenvolvimento do cálculo de variações:** Hilbert clamava por novas ideias no cálculo de variações. Muito trabalho foi feito; questão vaga demais para ser considerada resolvida.

8. Reimpresso como: Jacques Hadamard, *The Psychology of Invention in the Mathematical Field*, Dover, 1954.

2. Território dos primos: a conjectura de Goldbach

1. O algoritmo de Agrawal-Kayal-Saxena é o seguinte:

Input: *n* inteiro.

1. Se *n* for potência exata de qualquer número menor, responda COMPOSTO e pare.
2. Ache o menor *r* tal que a menor potência de *r* que seja igual a 1 módulo *n* seja pelo menos $(\ln n)^2$.
3. Se qualquer número menor ou igual a *r* tiver um fator em comum com *n*, responda composto e pare.
4. Se *n* for menor ou igual a *r*, responda PRIMO e pare.
5. Para todos os números inteiros *a* variando de 1 a um limite específico, verifique se o polinômio $(x + a)^n$ é o mesmo que $x^n + a$, para módulo *n* e para módulo $x^r - 1$. Se em qualquer caso a igualdade ocorrer, responda composto e pare.
6. Responda PRIMO.

2. Um exemplo do que tenho em mente é a fórmula $\lfloor A^{3^n} \rfloor$, onde os colchetes representam o maior inteiro menor ou igual ao seu conteúdo. Em 1947, W.H. Mills provou que existe uma constante real *A* tal que esta fórmula é prima para qualquer *n*. Assumindo a hipótese de Riemann, o menor valor de *A* que dá certo é em torno de 1.306. Entretanto, a constante é definida usando-se uma sequência adequada de primos, e a fórmula é simplesmente uma maneira simbólica de reproduzir essa sequência. Para mais fórmulas desse tipo, inclusive algumas que representam todos os primos, ver:

   http://mathworld.wolfram.com/PrimeFormula.html

   http://en.wikipedia.org/wiki/Formula_for_primes
3. Se *n* é ímpar, então $n - 3$ é par, e se *n* é maior que 5, então $n - 3$ é maior que 2. Pela primeira conjectura, $n - 3 = p + q$, logo, $n = p + q + 3$.
4. Também conhecida como "progressão aritmética".

3. O quebra-cabeça de pi: a quadratura do círculo

1. http://www.numberworld.org/misc_runs/pi-5t/details.html
2. Meu ódio predileto neste contexto é o termo "salto quântico". Em linguagem coloquial ele indica algum gigantesco salto adiante, ou alguma enorme mudança, como a descoberta da América pelos europeus. Em teoria quântica, porém, um salto quântico é tão minúsculo que nenhum instrumento conhecido pode observá-lo diretamente, uma mudança cujo tamanho é na casa de 0,000 … 01 com quarenta zeros ou mais.
3. Encontrar um dissecção finita de um quadrado num círculo é chamado problema da quadratura do círculo. Miklós Laczkovich o resolveu em 1990. Seu método é não construtivo e faz uso do axioma da escolha. O número de pedaços requerido é imenso, cerca de 10^{50}.
4. As bizarras alegações de quadradores de círculos e trissectores de ângulos são exploradas em profundidade em Dudley Underwood, *A Budget of Trisections*, Springer, 1987, e em *Mathematical Cranks*, da Mathematical Association of America, 1992. O fenômeno não é novo: ver Augustus De Morgan, *A Budget of Paradoxes*, Longmans, 1872; reimpresso pela Books for Libraries Press, 1915.
5. A quadratriz de Hípias é a curva traçada por uma reta vertical que se move uniformemente através de um retângulo e uma reta que gira uniformemente em torno do ponto médio da base desse retângulo (Figura 52). Essa relação transforma qualquer questão acerca de divisão de ângulos na questão correspondente sobre divisão de um segmento. Por exemplo, para trisseccionar um ângulo basta trisseccionar o segmento correspondente. Ver: http://www.geom.uiuc.edu/~huberty/math5337/groupe/quadratrix.html

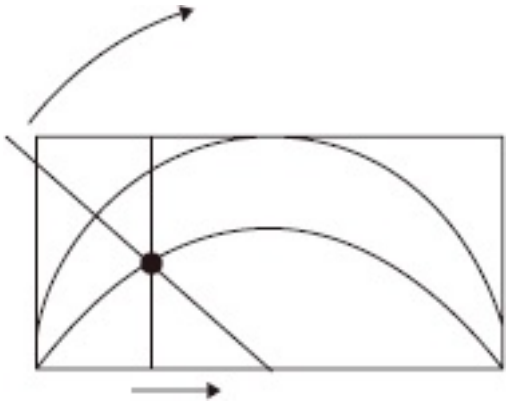


FIGURA 52 A quadratriz de Hípias (curva inferior).

6. Eis um exemplo explícito. Geometricamente, se uma reta cruza um círculo e não é tangente a ele, ela corta o círculo exatamente em dois pontos. Considere uma reta que seja paralela ao eixo horizontal, distância ½ acima dele (Figura 53). A equação dessa reta é muito simples: $y = ½$. (Qualquer que seja o valor de x, sempre teremos o mesmo valor para y.) Quando $y = ½$, a equação $x^2 + y^2 = 1$ torna-se $x^2 + ¼ = 1$. Portanto, $x^2 = {}^3/_4$, então $x = \sqrt{3}/2$ ou $-\sqrt{3}/2$. Logo, a álgebra nos diz que o círculo unitário encontra a nossa reta escolhida em exatamente dois pontos, cujas coordenadas são $(\sqrt{3}/2, ½)$ e $-(\sqrt{3}/2, ½)$. Isto é consistente com a Figura 53 e com o raciocínio puramente geométrico.

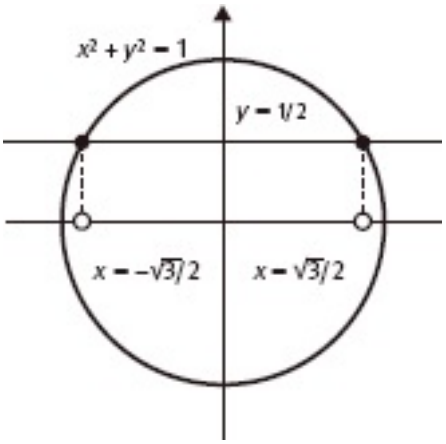


FIGURA 53 Uma reta horizontal cortando o círculo em dois pontos.

7. Estritamente falando, o polinômio em questão deve ter coeficientes inteiros e ser irredutível: não ser produto de dois polinômios de grau menor com coeficientes inteiros. Ter grau que seja potência de 2 nem sempre é suficiente para existir uma construção régua e compasso, mas é sempre necessário. Se o grau não é potência de 2, não existe construção. Se é potência de 2, é necessária análise adicional para decidir se a construção existe.
8. O inverso também é verdadeiro: dadas construções para polígonos regulares de três e cinco lados, pode-se deduzir a construção para o polígono de quinze lados. A ideia subjacente é que ${}^2/_5 - {}^1/_3 = {}^1/_{15}$. Um ponto sutil refere-se a potências primas. O argumento não fornece uma construção para, digamos, um eneágono, polígono de nove lados, dada por seus fatores primos – ou seja, triângulos. Gauss provou que nenhuma construção é possível para potências primas ímpares maiores que a primeira.
9. Ver Ian Stewart, *17 equações que mudaram o mundo*, Capítulo 5.
10. Para dar sentido a esta afirmação, resolve-se a equação quadrática em fatores lineares. Então $x^2 -$

$1 = (x + 1)(x - 1)$, que será zero se cada fator for zero, então $x = 1$ ou -1. O mesmo raciocínio aplica-se a $x^2 = xx$: será zero se ou o primeiro fator $x = 0$ ou o segundo fator $x = 0$. Acontece que essas duas soluções produzem o mesmo x, mas a ocorrência de dois fatores x distingue a situação de algo como $x(x - 1)$, onde há somente um fator x. Ao contar quantas soluções uma equação algébrica tem, a resposta é geralmente mais precisa se essas "multiplicidades" forem levadas em consideração.

11. Quando $n = 9$, o segundo fator é

    $x^8 + x^7 + x^6 + x^5 + x^4 + x^3 + x^2 + x + 1$

    Mas isso em si também tem fatores: é igual a

    $(x^2 + x + 1)\ (x^6 + x^3 + 1)$

    A caracterização de Gauss para números construtíveis requer que cada fator irredutível tenha grau que seja potência de 2. Mas o segundo fator tem grau 6, que não é potência de 2.

12. Gauss provou que o 17-ágono pode ser construído contanto que se consiga construir um segmento cujo comprimento seja:

    $$\frac{1}{16}\left[-1 + \sqrt{17} + \sqrt{34 - 2\sqrt{17}} + \sqrt{68 + 12\sqrt{17} - 16\sqrt{34 + 2\sqrt{17}} - 2(1 - \sqrt{17})(\sqrt{34 - 2\sqrt{17}})}\ \right]$$

    Como sempre é possível construir raízes quadradas, isso efetivamente soluciona o problema. Outros matemáticos encontraram construções explícitas. Ulrich von Huguenin publicou a primeira em 1803, e H.W. Richmond encontrou uma mais simples em 1893. Na Figura 54, pegue dois raios perpendiculares AOP_0 e BOC de um círculo. Faça OJ = ¼OB e o ângulo OJE = ¼OJP_0. Ache F de modo que o ângulo EJF seja de 45 graus. Desenhe um círculo com FP_0 como diâmetro, cruzando OB em K. Marque o centro do círculo E através de K, cortando AP_0 em G e H. Desenhe HP_3 e GP_5 perpendiculares a AP_0. Então P_0, P_3, P_5 são, respectivamente, os vértices 0, 3 e 5 de um 17-ágono regular, e os outros vértices podem ser agora facilmente construídos.

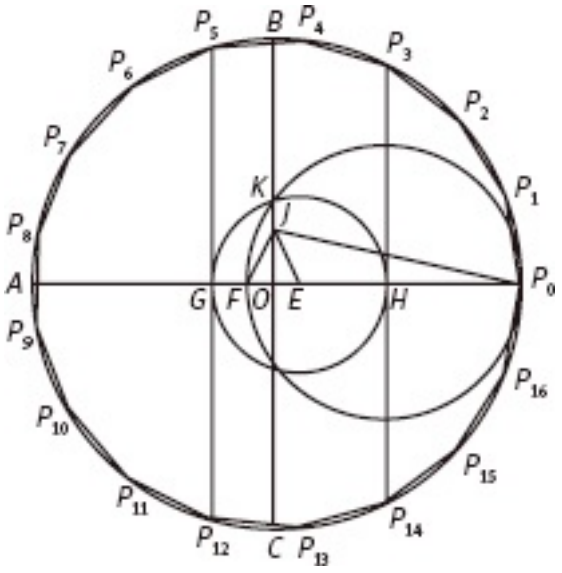


FIGURA 54 Como construir um 17-ágono regular.

13. Para as descobertas mais recentes, ver Wilfrid Keller, "Prime factors of Fermat numbers and complete factoring status", http://www.prothsearch.net/fermat.html
14. F.J. Richelot publicou uma construção para o 257-ágono regular em 1832. J. Hermes, da Universidade de Lingen, dedicou dez anos ao 65537-ágono. Seu trabalho, nunca publicado, pode

ser encontrado na Universidade de Göttingen, mas acredita-se que contenha erros.

15. Uma fração contínua típica tem o seguinte aspecto:

$$3 + \cfrac{1}{7 + \cfrac{1}{15 + \cfrac{1}{1 + \cfrac{1}{292 + \ldots}}}}$$

Esta fração contínua específica é o começo daquela que representa π.

16. http://bellard.org/pi-challenge/announce220997.html

4. Mistérios na elaboração de mapas: o teorema das quatro cores

1. Louis H. Kauffman, "Map coloring and the vector cross product", *Journal of Combinatorial Theory* B 48 (1990), p.145-54.
2. Se se permite que as fronteiras sejam realmente muito complicadas, não como um mapa, mas bem mais sinuosas, então quantos países você quiser poderão partilhar uma "divisa" comum. Uma construção chamada Lagos de Wada prova este resultado contraintuitivo. Ver: http://en.wikipedia.org/wiki/Lakes_of_Wada
3. Até recentemente, pensava-se que o artigo da *Nature* fosse a última referência impressa ao problema por quase um século, mas o historiador da matemática Robin Wilson descobriu este artigo posterior de Cayley.
4. Trabalhando no grafo dual, seja F o número de faces (incluindo uma face grande cercando todo o grafo), A o número de arestas (traços) e V o número de vértices. Podemos assumir que cada face no grafo dual tem pelo menos três arestas – se houver uma face com apenas duas arestas então ela corresponde a um vértice "supérfluo" do grafo original que junta apenas duas arestas. Este vértice pode ser apagado e as duas arestas, juntadas. Cada aresta é divisa de duas faces, e cada face tem pelo menos três arestas, então $A \geq {}^{3F}/_{2}$, ou equivalentemente ${}^{2A}/_{3} \geq F$. Pelo teorema de Euler, $F + V - A = 2$, então ${}^{2A}/_{3} + V - A \geq 2$, implicando

   $12 + 2A \leq 6V$

   Suponha que V_m seja o número de vértices com m vizinhos. Então V_2, V_3, V_4 e V_5 são zero. Portanto

   $V = V_6 + V_7 + V_8 + \ldots$

   Como toda aresta liga dois vértices

   $2A = 6V_6 + 7V_7 + 8V_8 + \ldots$

   Substituindo estes membros pela desigualdade, obtemos

   $12 + 6V_6 + 7V_7 + 8V_8 + \ldots \leq 6V_6 + 6V_7 + 6V_8 + \ldots$

   de modo que

   $12 + V_7 + 2V_8 + \ldots \leq 0$

   o que é impossível.

5. "Corrente" ou "cadeia" são termos enganosos, pois sugerem uma sequência linear. Uma corrente de Kempe contém laçadas e pode se ramificar.

6. A prova é dada por inteiro em Gerhard Ringel, *Map Color Theorem*, Springer, 1974. Ela faz a separação em doze casos, dependendo de o genus ser da forma $12k$, $12k + 1$, ..., $12k + 11$. Vamos chamá-los de casos 0-11. Com finitas exceções, os casos foram resolvidos como se segue:

   Caso 5: Ringel, em 1954.

   Casos 3, 7 e 10: Ringel, em 1961.

   Casos 0 e 4: C.M. Terry, Lloyd Welch e Youngs, em 1963.

   Caso 1: W. Gustin e Youngs, em 1964.

   Caso 9: Gustin, em 1965.

   Caso 6: Youngs, em 1966.

   Casos 2, 8 e 11: Ringel e Youngs, em 1967.

   As exceções foram genus 18, 20 e 23 (resolvidas por Yves Mayer em 1967) e 30, 35, 47 e 659 (resolvidas por Ringel e Youngs em 1968). Eles também lidaram com o problema análogo para superfícies unilaterais (como a faixa de Möbius, mas sem as bordas), que Heawood também havia abordado.

7. A impressionante história de como o *bug* foi descoberto,[g] e o que aconteceu depois, pode ser encontrada em http://en.wikipedia.org/wiki/Pentium_FDIV_bug

5. Simetria esférica: a conjectura de Kepler

1. Um site excelente para informação sobre a física dos flocos de neve é: http://www.its.caltech.edu/~atomic/snowcrystals/
2. C.A. Rogers, "The packing of equal spheres", in *Proceedings of the London Mathematical Society*, n.8, 1958, p.609-20.
3. Como o espaço é infinito, há infinitas esferas, então tanto o espaço como as esferas possuem um volume total infinito. Não podemos definir a densidade como sendo ∞/∞, porque esta razão não tem um valor numérico bem-definido. Em vez disso, consideramos as regiões cada vez maiores do espaço e adotamos o valor limite da proporção dessas regiões que as esferas preenchem.
4. http://hydra.nat.uni-magdeburg.de/packing/csq/csq49.html
5. C. Song, P. Wang e H.A. Makse, "A phase diagram for jammed matter", in *Nature*, n.453, 29 mai 2008, p.629-32.
6. Hai-Chau Chang e Lih-Chung Wang, "A simple proof of Thue's theorem on circle packing", in arXiv:1009.4322v1, 2010.
7. J.H. Lindsey, "Sphere packing in $\mathbb{R}^3$", *Mathematika*, n.33, 1986, p.137-47. D.J. Muder, "Putting the best face on a Voronoi polyhedron", in *Proceedings of the London Mathematical Society*, n.56, 1988, p.329-48.
8. Hales usou várias noções diferentes para o que estou chamando de gaiola. A final é "estrela de decomposição". Minha descrição omite algumas distinções cruciais no intuito de tornar a ideia básica compreensível.
9. Suponha que a região seja um polígono, como na Figura 55. Dado qualquer ponto que não esteja sobre a linha poligonal, existe uma linha reta deste ponto que saia de um círculo grande contendo o polígono, sem passar por nenhum vértice do polígono. (Há uma quantidade finita de vértices, mas infinitas retas para escolher.) Esta reta corta o polígono um número finito de vezes, e este número é ou par ou ímpar. Definimos o interior como consistindo em todos os pontos em que o número é ímpar e o exterior como consistindo em todos os pontos em que o número é par. Fica então simples provar que cada uma dessas regiões está interligada e que o polígono as separa.

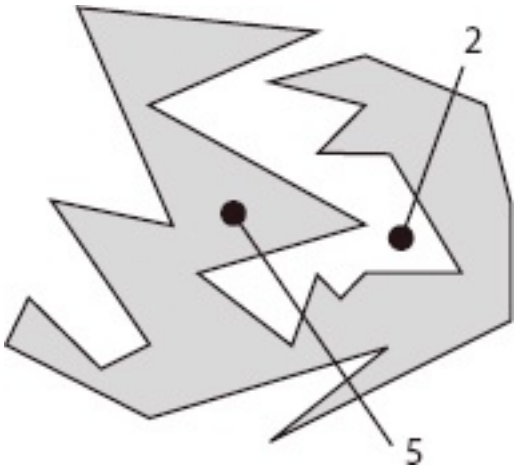


FIGURA 55 Prova do teorema da curva de Jordan para um polígono. Um número ímpar de intersecções ocorre para pontos na região sombreada (interior), e um número par de intersecções ocorre para os pontos da região clara (exterior).

10. http://code.google.com/p/flyspeck/

6. Novas soluções para coisas antigas: a conjectura de Mordell (p.130-44)

1. Andrew Granville e Thomas Tucker, "It's as easy as *abc*", in *Notices of the American Mathematical Society*, n.49, 2002, p.1224-31.

2. Para expandir este comentário críptico: a fórmula é

$$\int \frac{dx}{\sqrt{1-x^2}} = \operatorname{arcsen} x$$

onde arcsen (muitas vezes também escrito sen^{-1}) é a função inversa do seno. Isto é, se $y = \text{sen } x$, então $x = \text{arcsen } y$.

3. Por exemplo, seja k um número complexo, e considere a integral

$$\int \frac{dx}{\sqrt{(1-x^2)(1-k^2x^2)}}$$

Esta é a função inversa de uma função elíptica representada por sn. Há uma função dessas para cada valor de k. É o mesmo tipo de estrutura que o da nota 2, porém mais elaborada.

4. Ver Ian Stewart, *17 equações que mudaram o mundo*, Capítulo 8.

7. Margens inadequadas: o último teorema de Fermat

1. A prova pode ser encontrada em muitos textos de teoria dos números, por exemplo, Gareth A. Jones e J. Mary Jones, *Elementary Number Theory*, Springer, 1998, p.227. Na internet, ver: http://en.wikipedia.org/wiki/infinite_descent#Non-solvability_of_r2_.2B_s4_.3D_t4

2. A raiz de ordem p da unidade é o número complexo

$\zeta = \cos 2\pi/_p + i \operatorname{sen} 2\pi/_p$

e as outras são suas potências ζ^2, ζ^3, ..., ζ^{p-1}. Para ver por quê, lembrar que as funções trigonométricas seno e cosseno são definidas usando-se um triângulo retângulo, Figura 56 (esquerda). Para o ângulo A, usando os tradicionais a, b, c para os três lados, definimos o seno (sen) e o cosseno (cos) de A como

$\operatorname{sen} A = {}^{a}/_{c}$ $\cos A = {}^{b}/_{c}$

Se fizermos $c = 1$ e colocarmos o triângulo no plano complexo, como na Figura 56 (direita), o vértice no qual a e c se encontram é o ponto

$\cos A + i \operatorname{sen} A$

Agora é simples provar que para quaisquer ângulos A e B

$(\cos A + i \operatorname{sen} A)(\cos B + i \operatorname{sen} A) = \cos(A + B) + i \operatorname{sen}(A + B)$

e isto leva diretamente à fórmula de De Moivre

$(\cos A + i \operatorname{sen} A)^n = (\cos nA + i \operatorname{sen} nA)$

para qualquer inteiro positivo n. Portanto

$\zeta^p = (\cos 2\pi/_p + i \operatorname{sen} 2\pi/_p)^p = \cos 2\pi + i \operatorname{sen} 2\pi = 1$

logo, cada potência $1, \zeta, \zeta^2, \zeta^3, \ldots, \zeta^{p-1}$ é uma raiz de ordem p da unidade. Paramos aí porque $\zeta^p = 1$, então não surgirão números novos se pegarmos potências mais altas.

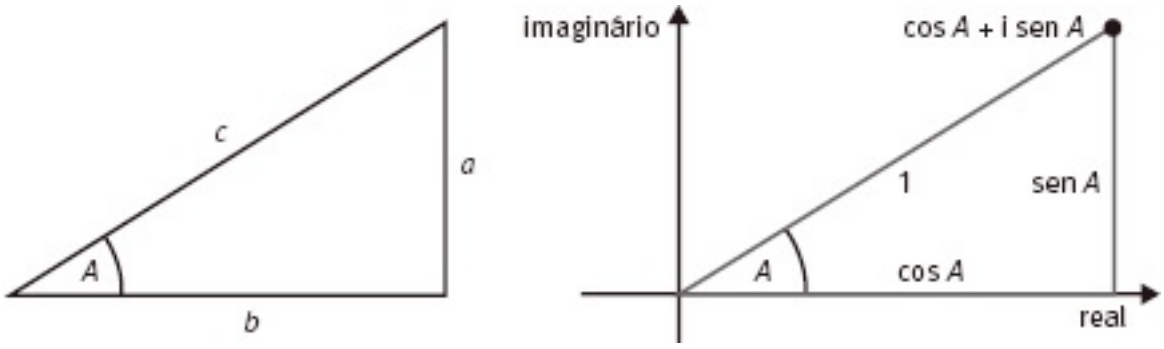


FIGURA 56 *Esquerda*: Definindo seno e cosseno.
Direita: Interpretação no plano complexo.

3. Apresentamos a *norma*

   $N(a + b\sqrt{15}) = a^2 - 15b^2$

   que tem a deliciosa propriedade

   $N(xy) = N(x)\,N(y)$

   Então

   $N(2) = 4$ $N(5) = 25$ $N(5+\sqrt{15}) = 10$ $N(5 + \sqrt{15}) = 10$

   Qualquer divisor próprio de um desses quatro números deve ter norma 2 ou 5 (divisores próprios de suas normas). Mas as equações $a^2 - 15b^2 = 2$ e $a^2 - 15b^2 = 5$ não têm soluções inteiras. Logo, não existem divisores próprios.

4. Simon Singh, *Fermat's Last Theorem*, Fourth Estate, 1997.

8. Caos orbital: o problema dos três corpos

1. Ou talvez não. Vladimir Krivchenkov demonstrou que a energia de estado básico e os primeiros estados excitados para o problema dos três corpos quântico pode ser calculada à mão. Mas em mecânica clássica, o problema análogo é menos tratável devido ao caos.
2. Citado em Arthur Koestler, *The Sleepwalkers*, Penguin Books, 1990, p.338.
3. Uma animação e informação adicional podem ser encontradas em http://www.scholarpedia.org/article/N-body_coreographies
4. Batizado em homenagem ao duque de Orrery, a quem foi apresentado em 1704.
5. Mais formalmente, é chamado de tempo de Liapunov.

9. Padrões em primos: a hipótese de Riemann

1. Há uma variante que integra $1/\ln t$ de 2 a x, em vez de 0 a x. Isso evita uma dificuldade técnica em $t = 0$, onde $\ln t$ não é definido. Às vezes a notação $\mathrm{Li}(x)$ é usada para esta variante, e a função definida no texto é chamada $\mathrm{li}(x)$.
2. O nome "Pafnuty" é incomum. Levou Philip Davis a escrever um livro peculiar, mas cativante: *The Thread: a Mathematical Yarn*, Harvester Press, 1983.
3. Isto é consequência da curiosa fórmula de Riemann

$$\zeta(1-s) = 2^{1-s}\pi^{-s}\,\mathrm{sen}\left(\frac{\pi(1-s)}{2}\right)\,\Gamma(s)\,\zeta(s)$$

   onde $\Gamma(s)$ é uma função clássica chamada função gama, definida para todo complexo s. O lado direito da equação é definido quando a parte real de s é maior que 1.
4. Bernhard Riemann, "Über die Anzahl der Primzahlen unter einer gegebenen Grösse", in *Monatsberichte der Königlich Preussichen Akademie der Wissenchaften zu Berlin*, nov 1859.
5. Riemann definiu uma função estreitamente ligada:

$$\Pi(x) = \pi(x) + \frac{1}{2}\pi\left(x^{1/2}\right) + \frac{1}{3}\pi\left(x^{1/3}\right) + \frac{1}{4}\pi\left(x^{1/4}\right) + \ldots$$

   que conta potências primas em vez de primos. Daí podemos recuperar $\pi(x)$. Então ele provou uma fórmula exata para esta função modificada em termos de integrais logarítmicas e uma integral correlacionada:

$$\Pi(x) = \mathrm{Li}(x) - \sum_{\rho}\mathrm{Li}(x^{\rho}) + \int_x^{\infty}\frac{\mathrm{d}t}{t(t^2-1)\ln t}$$

   Aqui $\sum$ indica uma soma de todos os números ρ para os quais $\zeta(\rho)$ é zero, excluindo inteiros pares negativos.

6. Por exemplo, $x + \sqrt{x}$ é assintótica a x: a razão é

$$\frac{(x+\sqrt{x})}{x} = \frac{1+1}{\sqrt{x}}$$

   À medida que x cresce, o mesmo ocorre com $\sqrt{x}$, então $1/\sqrt{x}$ tende a 0 e a razão tende a 1. Mas a diferença é $\sqrt{x}$, e esta torna-se cada vez maior à medida que x cresce. Por exemplo, quando x é 1 trilhão, $\sqrt{x}$ é 1 milhão.

7. A constante de Euler é o limite, quando n tende ao infinito, de

$$1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{n} - \ln n$$

8. Douglas A. Stoll e Patrick Demichel, "The impact of ζ(*s*) complex zeros on π(*x*) for $x < 10^{10^{13}}$, in *Mathematics of Computation*, n.276, 2011, p.2.381-94.
9. http://empslocal.ex.ac.uk/people/staff/mrwatkin/zeta/RHproofs.htm
10. J. Brian Conrey e Xian-Jin Li, "A note on some positivity conditions related to zeta- and *L*-functions": http://arxiv.org/abs/math.NT/9812166

10. Qual é o formato de uma esfera?: a conjectura de Poincaré

1. A 3-esfera unitária compreende todos os pontos com coordenadas (*x*, *y*, *z*, *w*) tais que $x^2 + y^2 + z^2 + w^2 = 1$. Há diversas maneiras de tornar a 3-esfera mais intuitiva. Elas podem ser todas entendidas por analogia com a 2-esfera e verificadas usando-se geometria de coordenadas. Uma descrição dessas ("bola sólida com toda a superfície identificada com um ponto") é dada no texto, e a Figura 57 mostra outra. Para estabelecer a analogia, observe que se cortarmos a 2-esfera ao longo de seu equador, obtemos duas hemiesferas. Cada uma delas achata-se num disco, e esta é uma deformação contínua. Para reconstruir a 2-esfera, simplesmente fazemos coincidir os pontos correspondentes nas bordas dos dois discos. Num sentido, fizemos um mapa da 2-esfera usando dois discos achatados, de forma muito parecida com a que os criadores de mapas fazem projeções planas do nosso planeta redondo. Podemos construir uma 3-esfera usando um procedimento análogo. Pegue duas bolas sólidas e faça coincidir pontos correspondentes sobre suas superfícies. Agora ambas têm a mesma superfície (porque as fizemos coincidir), e é uma 2-esfera. Ela forma o "equador" da 3-esfera.

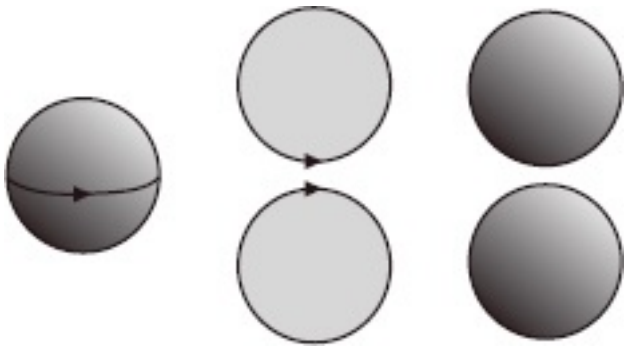


FIGURA 57 Como fazer uma 3-esfera. *Esquerda*: Cortar uma 2-esfera em hemiesferas. *Centro*: Reconstruir a 2-esfera a partir das duas metades colando as bordas. *Direita*: Por analogia, colar conceitualmente entre si as superfícies de duas bolas de modo que pontos correspondentes sejam considerados idênticos. Isso dá a 3-esfera.

2. A convenção usual é que falamos de adição e usamos a notação *a + b* quando a propriedade comutativa é válida, mas falamos de multiplicação e usamos a notação *ab* quando a propriedade pode não ser válida. Tenho ignorado essa convenção aqui porque este não é um livro-texto sobre teoria de grupos e "adição" parece mais natural.
3. Comece a contagem no zero. Toda vez que você passar pelo ponto de ônibus indo no sentido anti-horário, aumente 1 na contagem; toda vez que você passar pelo ponto no sentido horário, reduza 1 na contagem. No fim da viagem, some 1 se chegou pelo sentido anti-horário, subtraia 1 se chegou no sentido horário. A contagem final é o número de vezes que você deu a volta no círculo, medido no sentido anti-horário.

11. Não podem ser todos fáceis: o problema P/NP

1. A fórmula de Stirling afirma que $n!$ é aproximadamente $\sqrt{2\pi n}\ (n/e)^n$.
2. William J. Cook, *In Pursuit of the Travelling Salesman*, Princeton University Press, Princeton, 2012. Para informação atual, ver http://www.tsp.gatech.edu/index.html
3. Richard M. Karp, "Reducibility among combinatorial problems", in R.E. Miller e J.W. Thatcher (orgs.), *Complexity of Computer Computations*, Plenum, 1972, p.85-103.

12. Pensamento fluido: a equação de Navier-Stokes

1. Z. Xia, "The existence of noncollision singularities in Newtonian systems", in *Annals of Mathematics*, n.135, 1992, p.411-68.
2. http://www.claymath.org/millenium/Navier-Stokes_Equations/

13. Enigma quântico: a hipótese do *mass gap*

1. Ver Ian Stewart, *17 equações que mudaram o mundo*, Capítulo 14.

14. Sonhos diofantinos: a conjectura Birch–Swinnerton-Dyer

1. Leonardo Pisano Fibonacci, *The Book of Squares*, anotado e traduzido para o inglês por L.E. Sigler, Academic Press, 1987.
2. Leonardo encontrou uma família de soluções

$$\left(\frac{m^2+n^2}{2}\right)^2 - mn(m^2-n^2) = \left(\frac{m^2-2mn-n^2}{2}\right)^2$$

$$\left(\frac{m^2+n^2}{2}\right)^2 + mn(m^2-n^2) = \left(\frac{m^2+2mn-n^2}{2}\right)^2$$

   onde m, n são ambos ímpares. O papel de d aqui é desempenhado pelo número $mn(m^2 - n^2)$, e x é $m^2 + n^2/2$. Escolhendo $m = 5$, $n = 4$, chegamos a $mn(m^2 - n^2) = 720$. Temos que $720 = 5 \times 12^2$. Dividindo x por 12 obtém-se a resposta.
3. Se $x - n$, x e $x + n$ são quadrados, então também é seu produto, $x^3 - n^2x$. Portanto, a equação $y^2 = x^3 - n^2x$ tem uma solução racional. Além disso, y não é zero, senão $x = n$ e ambos x e $2x$ são quadrados, o que é impossível, uma vez que $\sqrt{2}$ é irracional.

   Inversamente, se x e y satisfazem a equação cúbica e y não é zero, então $a = (x^2 - n^2)/y$, $b = 2nx/y$ e $c = (x^2+n^2)/y$ satisfazem as equações $a^2 + b^2 = c^2$ e $ab/2 = n$.
4. Isto é,

$$\prod_{p \le x} \frac{N_p}{p} \approx C(\ln r)^r$$

   onde r é a ordem, C uma constante, e ≈ significa que a razão dos dois lados tende a 1 quando x tende ao infinito.

15. Ciclos complexos: a conjectura de Hodge

1. A razão mais provável é que estas sejam traduções naturais de linguagens usadas pelos matemáticos mais proeminentes nas duas áreas.
2. Por que *b* não era número de bananas, não tenho certeza. Quem sabe porque na Grã-Bretanha pós-guerra as bananas fossem itens exóticos raramente vistos nas lojas?
3. Daí uma piada padrão dos matemáticos. Um biólogo, um estatístico e um matemático estão sentados em um café observando o mundo passar. Um homem e uma mulher entram num prédio do outro lado da rua. Dez minutos depois, saem acompanhados por uma criança. "Eles reproduziram", diz o biólogo. "Não", diz o estatístico. "É um erro de observação. Em média, duas pessoas e meia foram em cada direção." "Não, não, não", diz o matemático. "É perfeitamente óbvio. Se alguém entrar agora, o prédio estará vazio."

16. E agora, para onde?

1. Bohr pode ter tido seriamente razão. Teorias científicas são testadas por meio de suas predições, mas poucas delas preveem o futuro. A maioria são afirmações do tipo se/então: se você passar luz através de um prisma ela se dividirá em cores. A "predição" não diz quando isso vai acontecer. Assim, paradoxalmente, podemos fazer previsões sobre o clima sem prever o clima. "Se o ar quente de um ciclone encontra o ar frio então vai nevar" é uma predição científica, mas não uma previsão do tempo.
2. A citação, ou uma variante próxima, tem sido atribuída a cerca de trinta fontes diferentes, inclusive Sam Goldwyn, Woody Allen, Winston Churchill e Confúcio. Ver: http://www.larry.denenberg.com/predictions.html

17. Doze para o futuro

1. Para as informações mais recentes, ver Prime Pages: http://primes.utm.edu
2. Ilia Krasikov e Jeffrey C. Lagarias, "Bounds for the $3x + 1$ problem using difference inequalities", in *Acta Arithmetica*, n.109, 2003, p.237-58.
3. Jorge F. Sawyer e Clifford A. Reiter, "Perfect parallelepipeds exist", 2009, in arXiv:0907.0220).
4. R. Fulek e J. Pach, "A computational approach to Conway's thrackle conjecture", in *Computational Geometry*, n.44, 2011, p.345-55.
5. http://en.wikipedia.org/wiki/Langton%27s_ant
6. A conjectura ABC afirma: Para qualquer $\varepsilon > 0$ existe uma constante $k_\varepsilon > 0$ tal que se a, b, c são inteiros positivos, não tendo fator comum maior que 1, e $a + b = c$, então $c \leq k_\varepsilon P^{1+\varepsilon}$, onde P é o produto de todos os primos distintos que dividem abc.
7. Andrew Granville e Thomas J. Tucker, "It's easy as *abc*", in *Notices of the American Mathematical Society*, n.49, 2002, p.1224-31.

   Em setembro de 2012 Shinichi Mochizuki anunciou que havia provado a conjectura ABC usando uma abordagem radicalmente nova aos fundamentos da geometria algébrica. Os estudiosos estão agora checando sua prova de quinhentas páginas, mas isso pode levar um bom tempo.

---

[g] A origem do termo "*bug*" – percevejo, pequeno besouro – para referir-se a problemas pequenos, porém complicados, em programas de computador provém da história mencionada pelo autor. (N.T.)